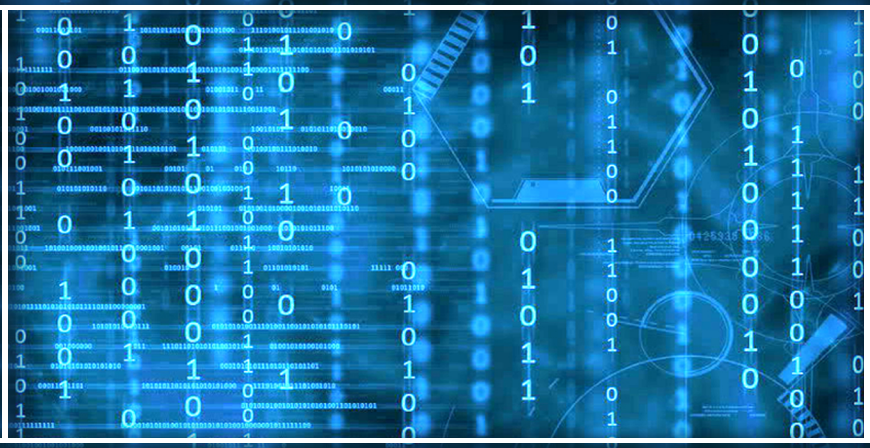
IJACSA
WHERE WISDOM SHARES
International Journal of Advanced Computer Science and Applications

Volume 7 Issue 11

November 2016



ISSN 2156-5570(Online)
ISSN 2158-107X(Print)

www.ijacsa.thesai.org

Editorial Preface

From the Desk of Managing Editor…

It may be difficult to imagine that almost half a century ago we used computers far less sophisticated than current home desktop computers to put a man on the moon. In that 50 year span, the field of computer science has exploded.

Computer science has opened new avenues for thought and experimentation. What began as a way to simplify the calculation process has given birth to technology once only imagined by the human mind. The ability to communicate and share ideas even though collaborators are half a world away and exploration of not just the stars above but the internal workings of the human genome are some of the ways that this field has moved at an exponential pace.

At the International Journal of Advanced Computer Science and Applications it is our mission to provide an outlet for quality research. We want to promote universal access and opportunities for the international scientific community to share and disseminate scientific and technical information.

We believe in spreading knowledge of computer science and its applications to all classes of audiences. That is why we deliver up-to-date, authoritative coverage and offer open access of all our articles. Our archives have served as a place to provoke philosophical, theoretical, and empirical ideas from some of the finest minds in the field.

We utilize the talents and experience of editor and reviewers working at Universities and Institutions from around the world. We would like to express our gratitude to all authors, whose research results have been published in our journal, as well as our referees for their in-depth evaluations. Our high standards are maintained through a double blind review process.

We hope that this edition of IJACSA inspires and entices you to submit your own contributions in upcoming issues. Thank you for sharing wisdom.

Thank you for Sharing Wisdom!

Managing Editor
IJACSA
Volume 7 Issue 11 November 2016
ISSN 2156-5570 (Online)
ISSN 2158-107X (Print)

Editorial Board

Reviewer Board Members

CONTENTS

Paper 1: BITRU: Binary Version of the NTRU Public Key Cryptosystem via Binary Algebra
Authors: Nadia M.G. Alsaidi, Hassan R. Yassein
PAGE 1 – 6

Paper 2: OWLMap: Fully Automatic Mapping of Ontology into Relational Database Schema
Authors: Humaira Afzal, Mahwish Waqas, Tabbassum Naz
PAGE 7 – 15

Paper 3: Vismarkmap – A Web Search Visualization Technique through Visual Bookmarking Approach with Mind Map Method
Authors: Abdullah Al-Mamun, Sheak Rashed Haider Noori
PAGE 16 – 23

Paper 4: A Sales Forecasting Model in Automotive Industry using Adaptive Neuro-Fuzzy Inference System(Anfis) and Genetic Algorithm(GA)
Authors: Amirmahmood Vahabi, Shahrooz Seyyedi Hosseininia, Mahmood Alborzi
PAGE 24 – 30

Paper 5: Japanese Dairy Cattle Productivity Analysis using Bayesian Network Model (BNM)
Authors: Iqbal Ahmed, Kenji Endo, Osamu Fukuda, Kohei Arai, Hiroshi Okumura, Kenichi Yamashita
PAGE 31 – 37

Paper 6: Analysis of Security Requirements Engineering: Towards a Comprehensive Approach
Authors: Ilham Maskani, Jaouad Boutahar, Souhaïl El Ghazi El Houssaïni
PAGE 38 – 45

Paper 7: Teachme, A Gesture Recognition System with Customization Feature
Authors: Hazem Qattous, Bilal Sowan, Omar AlSheikSalem
PAGE 46 – 50

Paper 8: A Mobile Device Software to Improve Construction Sites Communications "MoSIC"
Authors: Adel Khelifi, Khaled Hesham Hyari
PAGE 51 – 58

Paper 9: Framework of Resource Management using Server Consolidation to Minimize Live Migration and Load Balancing
Authors: Alexander Ngenzi, Selvarani R, Suchithra R
PAGE 59 – 64

Paper 10: Automatic Rotation Recovery Algorithm for Accurate Digital Image and Video Watermarks Extraction
Authors: Nasr addin Ahmed Salem Al-maweri, Aznul Qalid Md Sabri, Ali Mohammed Mansoor
PAGE 65 – 72

Paper 11: A Comparative Study Between the Capabilities of MySQl Vs. MongoDB as a Back-End for an Online Platform
Authors: Cornelia Győrödi, Robert Győrödi, Ioana Andrada Olah, Livia Bandici
PAGE 73 – 78

Paper 12: Security Risk Assessment of Cloud Computing Services in a Networked Environment
Authors: Eli WEINTRAUB, Yuval COHEN
PAGE 79 – 90

Paper 13: Using Multiple Seasonal Holt-Winters Exponential Smoothing to Predict Cloud Resource Provisioning
Authors: Ashraf A. Shahin
PAGE 91 – 96

Paper 14: Optimal Path Planning using RRT* based Approaches: A Survey and Future Directions
Authors: Iram Noreen, Amna Khan, Zulfiqar Habib
PAGE 97 – 107

Paper 15: Performance Analysis of In-Network Caching in Content-Centric Advanced Metering Infrastructure
Authors: Nour El Houda Ben Youssef, Yosra Barouni, Sofiane Khalfallah, Jaleleddine Ben Hadj Slama, Khaled Ben Driss
PAGE 108 – 115

Paper 16: Development of Dynamic Real-Time Navigation System
Authors: Shun FUJITA, Kayoko YAMAMOTO
PAGE 116 – 130

Paper 17: MIMC: Middleware for Identifying & Mitigating Congestion Level in Hybrid Mobile Adhoc Network
Authors: P. G. Sunitha Hiremath, C.V. Guru Rao
PAGE 131 – 139

Paper 18: Statistical Implicative Similarity Measures for User-based Collaborative Filtering Recommender System
Authors: Nghia Quoc Phan, Phuong Hoai Dang, Hiep Xuan Huynh
PAGE 140 – 146

Paper 19: Applying Chatbots to the Internet of Things: Opportunities and Architectural Elements
Authors: Rohan Kar, Rishin Haldar
PAGE 1 47– 154

Paper 20: State of the Art Exploration Systems for Linked Data: A Review
Authors: Karwan Jacksi, Nazife Dimililer, Subhi R. M. Zeebaree
PAGE 155 – 164

Paper 21: Qos-based Computing Resources Partitioning between Virtual Machines in the Cloud Architecture
Authors: Evgeny Nikulchev, Evgeniy Pluzhnik, Oleg Lukyanchikov, Dmitry Biryukov, Elena Andrianova
PAGE 165 – 170

Paper 22: Multiobjective Optimization for the Forecasting Models on the Base of the Strictly Binary Trees
Authors: Nadezhda Astakhova, Liliya Demidova, Evgeny Nikulchev
PAGE 171 – 179

Paper 23: Big Data Knowledge Mining
Authors: Huda Umar Banuqitah, Fathy Eassa, Kamal Jambi, Maysoon Abulkhair
PAGE 180 – 189

Paper 24: Characterizations of Flexible Wearable Antenna based on Rubber Substrate
Authors: Saadat Hanif Dar, Jameel Ahmed, Muhammad Raees
PAGE 190 – 195

Paper 25: E-Commerce Adoption at Customer Level in Jordan: an Empirical Study of Philadelphia General Supplies
Authors: Mohammed Al Masarweh, Sultan Al-Masaeed, Laila Al-Qaisi, Ziad Hunaiti
PAGE 196 – 205

Paper 26: Wavelet based Scalable Edge Detector
Authors: Imran Touqir, Adil Masood Siddique, Yasir Saleem
PAGE 206 – 211

Paper 27: Variability Management in Business-IT Alignment: MDA based Approach
Authors: Hanae Sbai, Mounia Fredj
PAGE 212 – 221

Paper 28: Performance Metrics for Decision Support in Big Data vs. Traditional RDBMS Tools & Technologies
Authors: Alazar Baharu, Durga Prasad Sharma
PAGE 222 – 228

Paper 29: Solving Word Tile Puzzle using Bee Colony Algorithm
Authors: Erum Naz, Khaled Al-Dabbas, Mahdi Abrishami, Lars Mehnen, Milan Cvetkovic
PAGE 229 – 234

Paper 30: A Novel Approach to Automatic Road-Accident Detection using Machine Vision Techniques
Authors: Vaishnavi Ravindran, Lavanya Viswanathan, Shanta Rangaswamy
PAGE 235 – 242

Paper 31: Real-Time Implementation of an Open-Circuit Dc-Bus Capacitor Fault Diagnosis Method for a Three-Level NPC Rectifier
Authors: Fatma Ezzahra LAHOUAR, Mahmoud HAMOUDA, Jaleleddine BEN HADJ SLAMA
PAGE 243 – 247

Paper 32: Issue Tracking System based on Ontology and Semantic Similarity Computation
Authors: Habes Alkhraisat
PAGE 248 – 251

Paper 33: Constraints in the IoT: The World in 2020 and Beyond
Authors: Asma Haroon, Munam Ali Shah, Yousra Asim, Wajeeha Naeem, Muhammad Kamran, Qaisar Javaid
PAGE 252 – 271

Paper 34: ETEEM- Extended Traffic Aware Energy Efficient MAC Scheme for WSNs
Authors: Younas Khan, Sheeraz Ahmed, Fakhri Alam Khan, Imran Ahmad, Saqib Shahid Rahim, M. Irfan Khattak
PAGE 272 – 277

Paper 35: Intelligent System for Detection of Abnormalities in Human Cancerous Cells and Tissues
Authors: Jamil Ahmed Chandio, M. Abdul Rahman Soomrani
PAGE 278 – 284

Paper 36: Enhanced Re-Engineering Mechnanism to Improve the Efficiency of Software Re-Engineering
Authors: A. Cathreen Graciamary, Chidambaram
PAGE 285 – 290

Paper 37: Scalable Scientific Workflows Management System SWFMS
Authors: M. Abdul Rahman
PAGE 291 – 296

Paper 38: Efficient Relay Selection Scheme based on Fuzzy Logic for Cooperative Communication
Authors: Shakeel Ahmad Waqas, Imran Touqir, Nasir Khan, Imran Rashid
PAGE 297 – 303

Paper 39: Wavelet-based Image Modelling for Compression Using Hidden Markov Model
Authors: Muhammad Usman Riaz, Imran Touqir, Maham Haider
PAGE 304 – 310

Paper 40: Image De-Noising and Compression Using Statistical based Thresholding in 2-D Discrete Wavelet Transform
Authors: Qazi Mazhar, Adil Masood Siddique, Imran Touqir, Adnan Ahmad Khan
PAGE 311 – 316

Paper 41: Denoising in Wavelet Domain Using Probabilistic Graphical Models
Authors: Maham Haider, Muhammad Usman Riaz, Imran Touqir, Adil Masood Siddiqui
PAGE 317 – 321

Paper 42: Connected Dominating Set based Optimized Routing Protocol for Wireless Sensor Networks
Authors: Hamza Faheem, Naveed Ilyas, Siraj ul Muneer, Sadaf Tanvir
PAGE 322 – 331

Paper 43: Adaptive Error Detection Method for P300-based Spelling Using Riemannian Geometry
Authors: Attaullah Sahito, M. Abdul Rahman, Jamil Ahmed
PAGE 332 – 337

Paper 44: Evaluation of OLSR Protocol Implementations using Analytical Hierarchical Process (AHP)
Authors: Ashfaq Ahmad Malik, Tariq Mairaj Rasool Khan, Athar Mahboob
PAGE 338 – 344

Paper 45: Fast Approximation for Toeplitz, Tridiagonal, Symmetric and Positive Definite Linear Systems that Grow Over Time
Authors: Pedro Mayorga, Alfonso Estudillo, A. Medina-Santiago, Jos´e V´azquez, Fernando Ramos
PAGE 345 – 350

Paper 46: A Multi-Agent Framework for Data Extraction,Transformation and Loading in Data Warehouse
Authors: Ramzan Talib, Muhammad Kashif Hanif, Fakeeha Fatima, Shaeela Ayesha
PAGE 351 – 354

Paper 47: Polynomial based Channel Estimation Technique with Sliding Window for M-QAM Systems
Authors: O. O. Ogundile, M. O. Oloyede, F. A. Aina, S. S. Oyewobi
PAGE 355 – 358

Paper 48: Synergies of Advanced Technologies and Role of VANET in Logistics and Transportation
Authors: Kishwer Abdul Khaliq, Amir Qayyum, Jurgen Pannek
PAGE 359 – 369

Paper 49: WQbZS: Wavelet Quantization by Z-Scores for JPEG2000
Authors: Jesus Jaime Moreno-Escobar, Oswaldo Morales-Matamoros, Ricardo Tejeida-Padilla, Ana Lilia Coria-Paes, Teresa Ivonne Contreras-Troya
PAGE 370 – 378

Paper 50: Determination of Child Vulnerability Level from a Decision-Making System based on a Probabilistic Model
Authors: SAHA Kouassi Bernard, BROU Konan Marcelin, Gooré Bi Tra, Souleymane OUMTANAGA
PAGE 379 – 384

Paper 51: Software-Defined Networks (SDNs) and Internet of Things (IoTs): A Qualitative Prediction for 2020
Authors: Sahrish Khan Tayyaba, Munam Ali Shah, Naila Sher Afzal Khan, Yousra Asim, Wajeeha Naeem, Muhammad Kamran
PAGE 385 – 404

Paper 52: Modified Random Forest Approach for Resource Allocation in 5G Network
Authors: Parnika De, Shailendra Singh
PAGE 405 – 413

Paper 53: Text Mining: Techniques, Applications and Issues
Authors: Ramzan Talib, Muhammad Kashif Hanif, Shaeela Ayesha, Fakeeha Fatima
PAGE 414 – 418

Paper 54: A Generic Model for Assessing Multilevel Security-Critical Object-Oriented Programs
Authors: Bandar M. Alshammari
PAGE 419 – 427

Paper 55: Towards Analytical Modeling for Persuasive Design Choices in Mobile Apps
Authors: Hamid Mukhtar
PAGE 428 – 434

Paper 56: Computer Science Approach TTo Philosophy: Schematizing Whitehead's Processes
Authors: Sabah Al-Fedaghi
PAGE 435 – 443

Paper 57: Mood Extraction Using Facial Features to Improve Learning Curves of Students in E-Learning Systems
Authors: Abdulkareem Al-Alwani
PAGE 444 – 453

Paper 58: Impact of Domain Modeling Techniques on the Quality of Domain Model: An Experiment
Authors: Hiqmat Nisa, Salma Imtiaz, Muhammad Uzair Khan, Saima Imtiaz
PAGE 454 – 462

BITRU: Binary Version of the NTRU Public Key Cryptosystem via Binary Algebra

Nadia M.G. Alsaidi
Department of applied Sciences
University of Technology
Baghdad, Iraq

Hassan R. Yassein
Department of Mathematics
College of Education, Al-Qadisiyah University
AL-Dewaniya, Iraq

***Abstract*—New terms such as closest vector problem (CVP) and the shortest vector problem (SVP), which have been illustrated as NP-hard problem, emerged, leading to a new hope for designing public key cryptosystem based on certain lattice hardness. A new cryptosystem called NTRU is proven computationally efficient and it can be implemented with low cost. With these characteristics, NTRU possesses advantage over others system that rely on number-theoretical problem in a finite field (e.g. integer factorization problem or discrete logarithm problem). These advantages make NTRU a good choice for many applications. After the adaptation of NTRU, many attempts to generalize its algebraic structure have appeared. In this study, a new variant of the NTRU public key cryptosystem called BITRU is proposed. BITRU is based on a new algebraic structure used as an alternative to NTRU-mathematical structure called binary algebra. This commutative and associative. Establishing two public keys in the proposed system has distinguished it from NTRU and those similar to NTRU cryptosystems. This new structure helps to increase the security and complexity of BITRU. The clauses of BITRU, which include key generation, encryption, decryption, and decryption failure, are explained in details. Its suitability of the proposed system is proven and its security is demonstrated by comparing it with NTRU.**

***Keywords*—NTRU; BITRU; polynomial ring; binary algebra**

I. INTRODUCTION

With the rapid development of wireless communication system widely deployed in recent years, security has become a crucial issue. Cryptography to solve this issue; it is used to meet the requirements of data and network communication security, namely, confidentiality, integrity, authentication, and non-repudiation [1]. The designing of high-performed algorithms is greatly demanded, which leads to security risk and heightens the need for analysis and investigation. Many public key cryptosystems have been developed since the Diffie Hellman seminal paper [2] was presented in 1976. Most of these cryptosystem are based on two mathematical hard problems: factorization and discrete logarithm problems (e.g., RSA [3], ElGamal cryptosystem [4], ECC [5], and many others [6]). From a practical perspective, most of these systems are costly because of their space complexity and high computation. This problem can be resolved by looking for new fast cryptosystems based on different hard problems.

The number theory research unit (NTRU) public key cryptosystem is a new generation of public key cryptosystems based on lattice hard problem introduced in 1996 by three mathematicians, namely Jeffery Hoffstein, Joseph Silverman, and Jill Piper [7]. It is the first public key cryptosystem that does not depend on the factorization and discrete algorithm problems aforementioned mathematical problems. Unfortunately, similar to many other public key systems, its security is unguaranteed although it is closely based on lattice problem. The basic collection of objects used by the NTRU public key cryptosystem occurs in a truncated polynomial ring of degree $N-1$ with integer coefficients belonging to $Z[x]/(x^N-1)$. NTRU is faster and has significantly smaller keys than the RSA and ECC cryptosystem.

Many researchers have improved the performance of NTRU by developing of its algebraic structure. In 2002, Gaborit et al. [8] introduced a NTRU-like cryptosystem called CTRU by replacing the base ring of the NTRU with a polynomial ring over a binary field F2[x]. They proved that their system is successfully decrypted. In 2005, Kouzmenko [9] showed that CTRU is weak under a time attack and proposed the GNTRU cryptosystem based on Gaussian integers Z[i] rather Z or $F_2[x]$. In the same year, Coglianese et al. [10] introduced an analog to the NTRU cryptosystem called MaTRU. MaTRU is based on a ring of all square matrices with polynomial entries. In 2009, Malekian et al. introduced the QTRU cryptosystem based on quaternion algebra [11]. They also introduced the OTRU cryptosystem in 2010 based on Octonion algebra [12]. Afterward, Vats [13] presented a new a non-commutative NTRU analog. His system is operated in the non-commutative ring

$M = M_k(Z)[X]/(X^n - I_{k\times k})$, where M is a matrix ring of the $k \times k$ matrices of polynomials in $Z[x]/(x^N-1)$. He proved that the speed is improved by a factor of $O(k^{1.624})$ over NTRU. In 2011, N. Zhao and S. Su [14] improved the algorithm of seeking the inverse of polynomial in NTRU. Also, they designed a new algorithm to judge whether the polynomial is invertible or not by computing $\gcd(\det(A), w)$. If it equals to 1, it is invertible, otherwise, the polynomial has no inverse in modulo w, and this algorithm use a matrix of an *N*–cyclic (A) corresponding to coefficients of polynomial of order N.

In 2012, Y. Bin Pan and Y. Deng in [15] focused on the technique of hiding the trapdoor of NTRU cryptosystem. So, they presented general NTRU – like framework. This framework has constructed new lattice based public key cryptosystem to find some particular kinds of easy closest vector problems (CVPs). They proposed a new lattice based public key cryptosystem as an application of their framework.

In 2013, Jarvis et al. [16] proposed a new framework based on the ring of a cubic root of unity known as the Eisenstein ring Z[w], whose coefficient integers belong to Z. They called it ETRU.

In 2014, P. Gauravaram, H. Narumanchi and N. Emmadi [17] present our analytical study on the implementation of NTRU encryption scheme which serves as a guideline for security practitioners who are novice to lattice based cryptographic implementations. In the same year, D. Cabarcas, P. Weiden, and J. Buchmann in [18] focused on the relationship between two embedding's ideals into geometric space and the shortest vector problem in principal ideal lattice.

In 2015, S. C. Batson in [19] focused on the relationship between two embedding's ideals into geometric space and the shortest vector problem in principal ideal lattice. In the same year, Alsaidi et al. [20] introduced the CQTRU cryptosystem based on commutative quaternion algebra.

In 2016, Thakur and Tripathi introduced BTRU, a new NTRU-like cryptosystem that replaces Z by a ring of polynomial with one variable over a rational field. They conveyed faster than NTRU [21]. In the same year, Yassein and Alsaidi [22] introduced an analog to the NTRU cryptosystem called HXDTRU, where the operations occur in the specially designed high-dimensional algebra called hexadecnion algebra.

In this study, we present a new multidimensional public key cryptosystem BITRU based on binary algebra. The mathematical structure of the proposed system results in two public keys, which in turn helps increase the BITRU security in comparison to its equivalents with identical structure.

This work is organized as follows. The summary of the original NTRU based on the arbitrary polynomial ring $Z[x]/(x^N - 1)$ is briefly introduced in Section II. The binary algebra used to construct the new NTRU-like cryptosystem, with its algebraic structure is provided in Section III. An analog of the NTRU cryptosystem called BITRU is proposed in Section IV. The successful decryption of the proposed system is proven through two propositions in Section V. The security and complexity analysis of the BITRU is discussed in Section VI. The study is concluded in Section VII.

II. NTRU Cryptosystem

A simple description of the NTRU cryptosystem is explained in this section. This cryptosystem depends on the addition and multiplication in the ring of a truncated polynomial of degree N denoted by $K = Z[X]/(X^N - 1)$, where N is a prime. Let $K_p(x) = (Z/pZ)[x]/(X^N - 1)$ and $K_q(x) = (Z/qZ)[x]/(X^N - 1)$denotes the rings of truncated polynomial modulo p and q respectively, where p and q are integers number, such that, $\gcd(p, q) = 1$ and q is significantly larger than p. Let d_f, d_g, d_m, and d_ϕ be constant integers less than N. Let L_f, L_g, L_m and $L_\phi \subset R$ be defined in Table 1.

TABLE I. Definition of the Public NTRU Parameters

Notation	Definition
L_f	{fϵ R | f has d_f coefficients equal to +1, ($d_f - 1$) equal to -1, the rest 0}
L_g	{gϵ R | g has d_g coefficients equal to +1, $d_g$equal to -1, the rest 0}
L_m	{mϵ R | coefficients of m are chosen modulo p, between $-p/2$ and $p/2$}
L_ϕ	{$\phi\epsilon$ R | ϕ has d_ϕ coefficients equal to +1, d_ϕ equal to -1, the rest 0}

A rough outline of the key creation, encryption, and decryption processes is presented as follows:

A. Key Generation

Public and private keys are generated by having the sender initially randomly choose two small polynomials f and g from L_f and L_g, respectively, such that f must be invertible modulo p and q denoted by F_p and F_q, respectively, where $f * F_p = 1$ and $f * F_q = 1$. A new polynomial f can be chosen if probable f is not invertible. Parameters f and g must be kept confidential. The public key h is computed in the following manner:

$$h = F_q * g \pmod q,$$

where f, F_p, F_q, and g are kept confidential (i.e., sender private key).

B. Encryption

Encryption is performed as follows:

For any given message $m \in L_m$, the public key h is used to compute the ciphertext e, such that,

$e = p\phi * h + m \pmod q$, where $\phi \in L_\phi$ is randomly chosen.

C. Decryption

Decryption is performed after the second party receives e. The receiver must find a, such that

$a = f * e \pmod q$, to derive the message. The coefficients of $a \epsilon K_q$ should be adjusted to lie in the interval $\left(-\frac{q}{2}, \frac{q}{2}\right]$, thus the unnecessary reduction of mod q.

$$\begin{aligned} a &= f * e \pmod q \\ &= f * (p\phi * h + m) \pmod q \\ &= pf * \phi * h + f * m \pmod q \\ &= pf * \phi * (F_q * g) + f * m \pmod q \\ &= p\phi * g + f * m \pmod q \end{aligned}$$

The resulting polynomial $p\phi * g + f * m$ obtains coefficients in the interval $(-q/2, q/2]$. It does not change if its coefficients are reduced to modulo q. The receiver computes the polynomial as follows:

$$b = a \pmod p$$

$$= p\phi * g + f * m \pmod p$$
$$= f * m \pmod p$$

The result is then multiplied by F_p to construct message m.

$$F_p * b = F_p * f * m \pmod p = m \pmod p,$$

the resulting coefficients are adjusted within the interval $[-q/2, q/2)$.

III. Binary Algebra

In this section, a real binary algebra and its properties are introduced. It is a vector space of two dimensions over the real numbers R defined as follows:

$BN_R = \{a + bj \,|a, b \in R\}$, where $j^2 = 1$ and R is the set of real numbers. The operation on this algebra is defined as follows:

Let $w_1, w_2 \in BN_R$, such that $w_1 = a_1 + b_1 j$ and $w_2 = a_2 + b_2 j$, the addition is then defined by

$w_1 + w_2 = (a_1 + a_2) + (b_1 + b_2)j$, the multiplication is then defined by

$w_1 . w_2 = (a_1 * a_2 + b_1 * b_2) + (a_1 * b_2 + b_1 * a_2)j$, and for any scalar r, the scalar multiplication is defined by $rw = ra + (rb)j$. This algebra is associative and commutative.

Every non zero element in BN_R $a + bj$ contains a unique multiplication inverse that is given by

$(a + bj)^{-1} = (\frac{b^2}{a(a^2-b^2)} + \frac{1}{a} - \frac{b}{a^2-b^2} j)$ such that $a^2 \neq b^2$.

Let F be a finite field of $\text{char}(F) \neq 2$. We define the binary algebra BN_F over F as follows: $BN_F = \{ a + bj \,|a, b \in F \}$, with addition, scalar multiplication, multiplication, and square norm as defined in the real binary algebra. We now consider the truncated polynomial ring

$K = Z[x]/(x^N - 1)$, $K_p(x) = (Z/pZ)[x]/(x^N 1)$ and $K_q(x) = (Z/qZ)[x]/(x^N - 1)$.

We define three binary algebras ψ, ψ_p, and ψ_q as follows:

$$\Psi = \{ f_0(x) + f_1(x)j \mid f_0, f_1 \in K \}$$
$$\Psi p = \{ f_0(x) + f_1(x)j \mid f_0, f_1 \in K_p \}$$
$$\Psi q = \{ f_0(x) + f_1(x)j \mid f_0, f_1 \in K_q \}.$$

Let ϕ_1 and $\phi_2 \in \psi_p$ or ψ_q, such that:

$$\phi_1 = f_0(x) + f_1(x)j$$
$$\phi_2 = g_0(x) + g_1(x)j,$$

where f_0, f_1 and $g_0, g_1 \in K_p$ or K_q.

The addition of ϕ_1 and ϕ_2 is performed by adding the corresponding coefficients mod p or mod q, such that $\phi_1 + \phi_2 = (f_0(x) + g_0(x)) + (f_1(x) + g_2(x))j$.

The multiplication of ϕ_1 and ϕ_2 is defined as follows:

$$\phi_1 * \phi_2 = (f_0 * g_0 + f_1 * g_1) + (f_0 * g_1 + f_1 * g_0)j,$$

where $*$ is the convolution product, the scalar multiplication is defined by $r\phi_1 = rf_0(x) + rf_1(x)j$ for any scalar r, and the same multiplication inverse is defined for the BN_R.

IV. Proposed BITRU Cryptosystem

The BITRU cryptosystem is set up by integers N, p, and q such that N is a prime, p and q are relatively prime and q is significantly larger than p. It also depends on five subsets define as follows

Definition 1: The subsets L_f, L_w, L_m, L_ϕ and $L_r \subset \Psi$ are called the subsets of BITRU defined as follows:

$L_f = \{f_0(x) + f_1(x)j \in \Psi \,|f_i(x)$ has d_f coefficients equal to 1, $d_f - 1$ equal to -1, the rest are $0\}$,

$L_w = \{w_0(x) + w_1(x)j \in \Psi | w_i(x)$ has d_w coefficients equal to 1, $d_w - 1$ equal to -1, the rest are $0\}$,

$L_m = \{m_0(x) + m_1(x)j \in \Psi \,|m_i(x)$ are chosen modulo p, between $-p/2$ and $p/2\}$,

$L_\phi = \{\phi_0(x) + \phi_1(x)j \in \Psi \,|\phi_i(x)$ has d_ϕ coefficients equal to 1, d_ϕ equal to -1, the rest are $0\}$ and

$L_r = \{r_0(x) + r_1(x)j \in \Psi \,|r_i(x)$ has d_r coefficients equal to +1, d_r equal to -1, the rest are $0\}$,

where d_f, d_w, d_ϕ and d_r are also constant parameters similar to those defined in the NTRU.

The BITRU cryptosystem is introduced based on the binary algebra and defined through four main phases described as follows:

A. Key Generation

The public and private keys are generated by making the sender randomly choose $f, g \in L_f, w \in L_w$, and $\phi \in L_\phi$, such that, f and g must have multiplicative inverse modulo p and q denoted by f_p, f_q and g_p, g_q respectively, and w have multiplicative inverse modulo p denoted by w_p.

The public keys are computed as follows:

$h = \phi f_q \pmod q$ ……. (1)
$k = g_q w \bmod (q)$ ……. (2)

where f, g, ϕ, and w are the private keys.

Algorithm 1 is designed for generating the first key set $h = [h_0, h_1]$

Algorithm 1: Ceatekey $[h_0, h_1]$

Input: $p, q, n, f_0, f_1, g_0, g_1$

1- $[Fq, Fq1] = \text{bininvq}\ (p, q, n, f_0, f_1)$
2- $[c_0, c_1] = \text{multbin}\ (g_0, g_1, Fq, F1, n, p)$
3- for i=1 to n
4- if $c_0(i) < 0$
5- $c_0(i) = c_0(i) + q$
6- end if
7- $c_0(i) = c_0(i) * p \pmod q$
8- end for
9- $h_0 = c_0$
10- for i = 1 to n

```
11-    if c1(i) < 0
12-      c1(i) = c1(i) + q
13-    end if
14-    c1(i) = c1(i) * p(mod q)
15- end for
16- h1 = c1
```

Algorithm 2 that is designed for generating of the second set $k = [k_0, k_1]$

Algorithm 2: generatekey $[k_0, k_1]$

Input: $p, q, n, f_0, f_1, g_0, g_1$

```
1-  [Fq, Fq1] = bininvq (p, q, n, f0, f1)
2-  [c0, c1] = multbin (g0, g1, Fq, F1, n, p)
3-  for i=1 to n
4-      if c0(i) < 0
5-        c0(i) = c0(i) + q
6-      end if
7-      c0(i) = c0(i) * p(mod q)
8-  end for
9-  k0 = c0
10- for i = 1 to n
11-    if c1(i) < 0
12-      c1(i) = c1(i) + q
13-    end if
14-    c1(i) = c1(i) * p(mod q)
15- end for
16- k1 = c1
```

B. Encryption

At the beginning of encryption, message m is converted to the binary algebra form, such that $m = m_0(x) + m_1(x)i$, where $m_i(x) \epsilon L_m$.

We choose $r \epsilon L_r$ which required the blinding value to encrypt the message $m \epsilon L_m$:

$$e = pr * h + m * k \,(\text{mod } q) \qquad \ldots\ldots\ldots (3)$$

Algorithm 1 is designed for encryption process

Algorithm 3: encryp $[e_0, e_1]$

Input: $n, m, q, m_0, m_1, h_0, h_1, k_0, k_1, r_0, r_1$

```
1- x = multbin(r0, r1, h0, h1, n, m)
2- y = multbin (m0, m1, k0, k1, n, m)
3- e0 = x mod q
4- e1 = y mod q
```

C. Decryption

After receiving e, it is left-multiplied by g and right-multiplied by f. Therefore,

$$a = g * e * f (\text{mod } q) \qquad \ldots\ldots\ldots (4)$$

where the coefficients of the polynomial a lie in the interval of $(-q/2 \text{ to } q/2]$.

$$b = a \,(\text{mod } p)$$

$$= w * m * f \,(\text{mod } p)$$

Compute $d = w_p * b * f_p \,(\text{mod } p)$.

Algorithm 4 that is designed for decryption

Algorithm 4: decryp $[d_0, d_1]$

Input: $n, p, q, f_0, f_1, g_0, g_1, w_0, w_1, e_0, e_1$

```
1-  [u0, u1] = multbin(g0, g1, e0, e1, n, q)
2-  [v0, v1] = multbin (u0, u1, f0, f1, n, q)
3-  [Fp0, Fp1] = bininvp (f0, f1, n, q)
4-  [wp0, wp1] = bininvp (w0, w1, n, q)
5-  for i=1 to n
6-      if v0(i) < 0
7-        v0(i) = v0(i) + q
8-      end if
9-    if v0(i) > (q/2)
10-     v0(i) = v0(i) − q
11-   end if
12- end for
13- for i=1 to n
14-     if v1(i) < 0
15-       v1(i) = v1(i) + q
16-     end if
17-     if v1(i) > (q/2)
18-       v1(i) = v1(i) − q
19-     end if
20- end for
21- [s0, s1] = multbin (w0, w1, v0, v1n, p)
22- [t0, t1] = multbin (s0, s1, Fp0, Fp1, n, p)
23- d0 = t0, d1 = t1
```

V. SUCCESSFUL DECRYPTION

Proposition: The polynomial d is computed by the receiver, and it is equal to the sender plaintext m.

Proof: $a = g * e * f \,(\text{mod } q)$

$$= g(pr * h + k * m) f \,(\text{mod } q) \qquad \text{from (3)}$$

$$= pg * r * h * f + g * k * m * f \,(\text{mod } q)$$

$$= pg * r * \phi * f_q * f + g * g_q * w * m * f \,(\text{mod } q)$$

from (1) and (2)

$$= pg * r * \phi + w * m * f \,(\text{mod } q).$$

Let $b = a \,(\text{mod } p)$

$$= pg * r * \phi + w * m * f \,(\text{mod } p).$$

The first term is equal to zero modulo p because it contains p.

$$b = w * m * f \,(\text{mod } p).$$

Then $d = w_p * b * f_p \,(\text{mod } p)$.

$$= w_p * w * m * f * f_p \,(\text{mod } p)$$

$$= m \,(\text{mod } p). \qquad \square$$

VI. LATTICE-BASED ATTACKS

To prove the security of BITRU, different attacks have been investigated to show that they are without major effects. In such cryptosystems that based on polynomial ring, the lattice is defined from the relation between the public key and the private key, where the private key represents the shortest vector in this lattice and can be found by solving the approximate matrix for that vector. The attacker must recover the private keys f and g from the public keys h and k, respectively, to attack BITRU. This move is equivalent to finding the shortest vector in the BITRU lattice denoted by $\mathcal{L}_{\mathrm{BITRU}}$.

The attacker first spreads $hf = \phi \pmod q$, $gk = w \pmod q$ as follows:

$$h_0 * f_0 + h_1 * f_1 = \phi_0 + qu_0$$
$$h_0 * f_1 + h_1 * f_0 = \phi_1 + qu_1$$

and

$$g_0 * k_0 + g_1 * k_1 = w_0 + qv_0$$
$$g_0 * k_1 + g_1 * k_0 = w_1 + qv_1.$$

All the polynomials h_0, h_1 and k_0, k_1 can be represented in their matrix isomorphic representation as follows:

$$(H_i)_{N\times N}=\begin{bmatrix} h_{j,0} & h_{j,1} & \dots & h_{j,N-1} \\ h_{j,N-1} & h_{j,0} & \dots & h_{j,N-2} \\ h_{j,N-2} & h_{j,N-1} & \dots & h_{j,N-3} \\ \vdots & & \ddots & \vdots \\ h_{j,2} & h_{j,3} & \dots & h_{j,1} \\ h_{j,1} & h_{j,2} & \dots & h_{j,0} \end{bmatrix}$$

and

$$(K_i)_{N\times N}=\begin{bmatrix} k_{j,0} & k_{j,1} & \dots & k_{j,N-1} \\ k_{j,N-1} & k_{j,0} & \dots & k_{j,N-2} \\ k_{j,N-2} & k_{j,N-1} & \dots & k_{j,N-3} \\ \vdots & & \ddots & \vdots \\ k_{j,2} & k_{j,3} & \dots & k_{j,1} \\ k_{j,1} & k_{j,2} & \dots & k_{j,0} \end{bmatrix} \quad j = 0,1$$

Therefore, $\mathcal{L}_{\mathrm{BITRU}}$ represented by $\mathcal{L}^{h}_{\mathrm{BITRU}}$ and $\mathcal{L}^{k}_{\mathrm{BITRU}}$ of dimension 8N are spanned by the rows of matrices

$$\mathcal{M}^{h}_{4N\times 4N} = \begin{bmatrix} I_{2Nx2N} & H_{2Nx2N} \\ 0_{2Nx2N} & qI_{2Nx2N} \end{bmatrix}$$

and

$$\mathcal{M}^{k}_{4N\times 4N} = \begin{bmatrix} I_{2Nx2N} & K_{2Nx2N} \\ 0_{2Nx2N} & qI_{2Nx2N} \end{bmatrix} \text{ respectively,}$$

where I denoted the identity matrix, qI denotes q times the identity matrix, 0 denotes zero matrix, and H, K are described as follows:

$$H_{2Nx2N} = \begin{bmatrix} h_0 & h_1 \\ -h_1 & -h_0 \end{bmatrix}$$

$$K_{2Nx2N} = \begin{bmatrix} k_0 & k_1 \\ -k_1 & -k_0 \end{bmatrix}$$

Therefore, the vectors $(\phi_0, \phi_1, f_0, f_1)$ and (g_0, g_1, w_0, w_1) belong to $\mathcal{L}^{h}_{\mathrm{BITRU}}$ and $\mathcal{L}^{k}_{\mathrm{BITRU}}$, respectively. A short vector in $\mathcal{L}^{h}_{\mathrm{BITRU}}$ and $\mathcal{L}^{k}_{\mathrm{BITRU}}$ can be found by a lattice reduction algorithm, which demonstrates that BITRU can resist lattice attacks significantly more than the NTRU. For simplicity, we assume that $d = df = d\phi = dw = dr \approx N/3$ because the determinant $\mathcal{L}^{h}_{\mathrm{BITRU}}$ is equal to the determinant of $\mathcal{M}^{h}_{4N\times 4N}$ which is an upper triangle matrix, and that its determinant is equal to q^{2N}, $\|(\phi_0, \phi_1, f_0, f_1)\| \approx \sqrt{8d} \approx 1.63\sqrt{N}$. The Gaussian heuristic expected that the length of the shortest nonzero vector is calculated as $\delta(\mathcal{L}^{h}_{\mathrm{BITRU}}) = \sqrt{\frac{2N}{\pi e}}\sqrt{q} \approx 0.48\sqrt{Nq}$. Also $\frac{\|(\phi_0,\phi_1,f_0,f_1)\|}{\delta} = \frac{1.63\sqrt{N}}{0.48\sqrt{Nq}} \approx \frac{3.39}{\sqrt{q}}$, hence the purpose vectors in $\mathcal{L}^{h}_{\mathrm{BITRU}}$ are shorter than that expected by the Gaussian heuristic, also the dimension of $\mathcal{L}^{h}_{\mathrm{BITRU}}$ is twice the time of the dimension of $\mathcal{L}_{\mathrm{NTRU}}$ when choosing the same value of N. In similar way, the length of the shortest nonzero vector is calculated as $\delta(\mathcal{L}^{k}_{\mathrm{BITRU}}) \approx 0.48\sqrt{Nq}$. Therefore, BITRU is more resistance against lattice attacks than NTRU.

VII. CONCLUSION

- In NTRU, the computation with small coefficient in the convolution product of polynomials resulted in a fast and low cost system that is superior to other theoretical number cryptosystems (e.g., RSA, ECC, and ElGamal) requiring a series of multiplications. The computation in NRTU also does not require any multi-precision libraries because all the polynomial coefficients are reduced mode q which resulted in 11 bit integers at most.
- In this study, the BITRU cryptosystem based on binary algebra is proposed. It is a multi-dimensional cryptosystem that can encrypt two messages from a single origin or two independent messages from two different origins. This property is important in certain applications such as, cellular phones and electronic voting system. When the coefficient of j is equal to zero.
- BITRU is converted to NTRU, with public key k=1 and g=1.
- The security of BITRU is four times that of NTRU because it contains two public keys h,k with four polynomials private keys f_0, f_1, g_0, and g_1.
- The proposed BITRU is a promising high-performing system. It exhibits certain robustness against well-known attacks that can threaten the security of the NTRU or NTRU-like cryptosystems.
- By lowering N, the speed of BITRU is faster than that of NTRU with the same parameters.

REFERENCES

[1] D. Robling, "Cryptography and data security," Addison – Wisely Publishing Company, 1982.

[2] W.Diffie, M. Hellman, "New directions in cryptography," IEEE Transactions On information theory, vol. 22, no.6, pp.644-654, 1976.

[3] R. Rivest, A. Shamir, L. Adleman, "A method for obtaining digital signature and public key cryptosystems, " Communications of the ACM, vol. 21, no. 2, pp.120-126, 1978.

[4] T. ElGamal, "A public key cryptosystem and a signature scheme based on discrete logarithm, " IEEE Transactions on Information Theory, vol. 31, no. 4, pp. 469-472, 1985.

[5] R.Schoof, "Elliptic curve over finite fields and the computation of square roots mod p, " Mathematics of computation, vol.44, no.170, pp. 483-494, 1985.

[6] R. McEliece, "A public key cryptosystem based on algebra coding theory," Pasadena, DSN Progress Reports 42-44, pp. 114-116, 1978.

[7] J. Hoffstein, J. Pipher, and J. Silverman, "NTRU: A ring based public key cryptosystem," Proceeding of ANTS III, LNCS, Springer Verlag, vol.1423, pp. 267-288, 1998.

[8] P. Gaborit J. Ohler, P. Soli, "CTRU, a polynomial analogue of NTRU," INRIA. Rapport de recherche, N. 4621, 2002.

[9] R. Kouzmenko, "Generalizations of the NTRU cryptosystem," Diploma Project, Ecole Polytechnique Federale de Lausanne, 2006.

[10] M. Coglianese and B. Goi, "MaTRU: A new NTRU based cryptosystem," Springer Verlag Berlin Heidelberg, pp. 232-243, 2005.

[11] E. Malecian, A. Zakerolhsooeini , A. Mashatan , "QTRU: a lattice attack resistant version of NTRU PCKS based on quaternion algebra," The ISC Int'l Journal of Information Security, vol. 3, no. 1, pp. 29–42, 2011.

[12] E. Malecian, A. Zakerolhsooeini, "OTRU: A non- associative and high speed public key cryptosystem," IEEE Computer Society, pp.83-90, 2010.

[13] N. Vats, NNRU a non-commutative analogue of NTRU, "CoRR, abs/0902.1891, 2009.

[14] N. Zhao and S. Su, "An improvement and a new design of Algorithms for Seeking the Inverse of NTRU polynomial", EEE Computer Society, Washington, 2011.

[15] Y. Pan and Y. Deng, "A General NTRU-Like Framework for Constructing Lattice Based Public Key Cryptosystems ", Springer-Verlag Berlin Heidelberg, p.p. 109–120, 2012.

[16] K. Jarvis and M. Nevins, "ETRU: NTRU over the Eisenstein integers," Springer Science +Business Media New York, 2013.

[17] P. Gauravaram, H. Narumanchi and N. Emmadi, " Analytical study of Implementation issues of NTRU", International Conference on Advances in Computing, Communications and Informatics, IEEE, New Delhi, India, pp. 700-707, 2014.

[18] S. C. Batson, "On the Relationship between Two Embeddings of Ideals into Geometric Space and the Shortest Vector Problem in Principal Ideal Lattices" Ph.D. thesis, North Carolina State University, 2015.

[19] N. Alsaidi, M. Said, A. Sadiq and A. Majeed, "An improved NTRU cryptosystem via commutative quaternions algebra," Int. Conf. Security and Management, SAM'15, pp.198-203, 2015.

[20] N. M. G. AlSaidi, M. Said, A. T. Sadiq, and A.A. Majeed, "An improved NTRU cryptosystem via commutative quaternions algebra," Int. Conf. Security and Management SAM'15, 2015, pp.198-203.

[21] K.Thakur and B.P.Tripathi, "BTRU, A Rational Polynomial Analogue of NTRU Cryptosystem," International Journal of Computer Applications, Foundation of Computer Science (FCS), NY, USA, vol. 145, no.12, 2016.

[22] H.R.Yassein, and N.AlSaidi, "HXDTRU Cryptosystem Based On Hexadecnion Algebra," 5th International Cryptology and Information Security Conference, 2016.

OWLMap: Fully Automatic Mapping of Ontology into Relational Database Schema

Humaira Afzal
Department of Computer Science
COMSATS Institute of Information Technology,
Lahore, Pakistan

Mahwish Waqas
Department of Computer Science
COMSATS Institute of Information Technology,
Lahore, Pakistan

Dr. Tabbassum Naz
Department of Computer Science & IT
University of Lahore,
Lahore, Pakistan

***Abstract*—Semantic web is becoming a controversial issue in current research era. There must be an automated approach to transform ontology constructs into relational database so that it can be queried efficiently. The previous research work based on transformation of RDF/OWL concepts into relational database contains flaws in complete transformation of ontology constructs into relational database. Some researchers claim that their technique of transformation is entirely automated, however their approach of mapping is incomplete and miss essential OWL constructs. This paper presents a tool called OWLMap that is fully automatic and provides lossless approach for transformation of ontology into relational database format. Number of experiments have been performed for ontology to relational database transformation. Experiments show that proposed approach is fully automatic, effective and quick. Our OWLMap is based on an approach that is lossless as well as it does not loose data, data types and structure.**

Keywords—Semantic Web; Ontology; Database; Mapping; OWL; Jena API

I. INTRODUCTION

The concept of semantic web is the extension of current web from human readable form to machine processable form by adding semantics. By applying structured information in semantic web, machines are capable to search, process, integrate and present the information in a meaningful and intelligent manner. Conventional search engines dissatisfy users by retrieving inadequate and inconsistent results because they work on predefined standards, terms that work in centralized environment. By semantic and ontology users are able to develop new facts and use their own keywords in different environment [1]. There are different techniques for storing ontology. Ontology can be stored in flat files [2]. But this technique does not provide scalability, query and other functionalities that database system can provide. Ontology repositories are used to hold ontology saved by Ontology management system [3]. But query facility in ontology management system is not as efficient as in relational database system. Relational database system has many advantages as compared to ontology management system like performance, robustness, maturity, reliability and availability. If ontology is stored in relational format then it can easily interoperate with large amount of existing web data. By using SQL, it is easy to retrieve information provided by ontology. If ontology is transformed into relational database then it will make semantic web more useful.

Previous researchers have worked on mapping of RDF/OWL concepts into relational database. But these mapping approaches have certain problems like loss of structure, loss of data and perform only initial mappings i.e. tables to classes and columns to properties. Most of transformation tools are semi-automatic and need human intervention [4].

We attempt to explain these problems and provide a solution in the form of OWLMap. The structure of this paper is as follows. In Section 2, previous approaches and their drawbacks have been provided. Section 3, describes proposed methodology. In Section 4, explains the implementation of our approach with the help of case study. Section 5; describe the important phase of testing. Section 6, concludes the main points of paper and give some future directions.

II. RELATED WORK

Reference [5] has purposed an approach for transformation of OWL to ER and vice versa by using conceptual graphs. The transformation is performed step by step, where the first phase is to transform the OWL ontology to ER and second phase is to transform ER to relational database. Reference [6] used "Oracle Semantic data storage" approach for transformation, but most OWL constructs are missing in this approach. Reference [7] suggested the "Storing ontology includes fuzzy data types" approach. Reference [8] purposed "large scale ontology management" approach that covers some constructs of OWL and transformation tool is not fully automatic.

Rule based transformation presented by [9] and [10] are based on "mapping rules". The short comes of these approach is that few constructs are missed during transformation. Few sub-properties and few constructs of OWL ontology are not considered e.g. property restrictions. OWL2DB algorithm is another approach to map OWL documents into relational tables without any human intervention [11]. The transformation is incomplete and it only saves class instances in relational format. Reference [12] proposed an approach for mapping of ontology to relational database. This approach is tested on ontology selected from product configuration domain. This approach covers only a few part of OWL DL syntax. In "Mapping of OWL ontology concepts to RDB Schemas" approach purposed by [13], authors purposed some mapping principles and algorithm. The prototype tool has been added as plug-in for an ontology editor named protégé.

It lacks some mappings like intersection, class complements, union, and property relations.

Reference [4] provides the state of the art for tools in the domain of automatic mapping of ontology into relational databases and highlights the need of fully automatic tool for transformation of ontology in to relational database.

Reference [14] purposed that it is required to have machine learning techniques for semantic mappings. Reference [15] developed a tool named "OntoRel" for transformation. The disadvantage of "OntoRel" was that it only selects few main OWL constructs for transformations. Reference [16] proposed a hybrid approach for reversible and lossless transformation. To improve query capabilities of the thus approach, more research is required.

III. METHODOLOGY

We have proposed a tool called OWLMap that is fully automatic in mapping ontology (OWL) to relational database format. In proposed system for transforming ontology to relational database format, initially a user will select an ontology file, and then information will be extracted about ontology constructs. After extracting this information, proposed mapping rules (given in section C) will be applied automatically to ensure lossless transformation.

A. System Architecture

In suggested approach for automatically convert OWL ontology to relational database format. Fig 1, explains an approach for transforming ontology to relational database format. As the figure depicts, initially, select an ontology file, then information is extracted about ontology constructs using Jena API. Mapping rules are defined to ensure lossless transformation. Based on the mapping rules, transformation of ontological constructs into relational database takes place. The main focus is to develop a tool from OWL ontology to relational database that is fully-automatic and can solve various problems from the previous approaches.

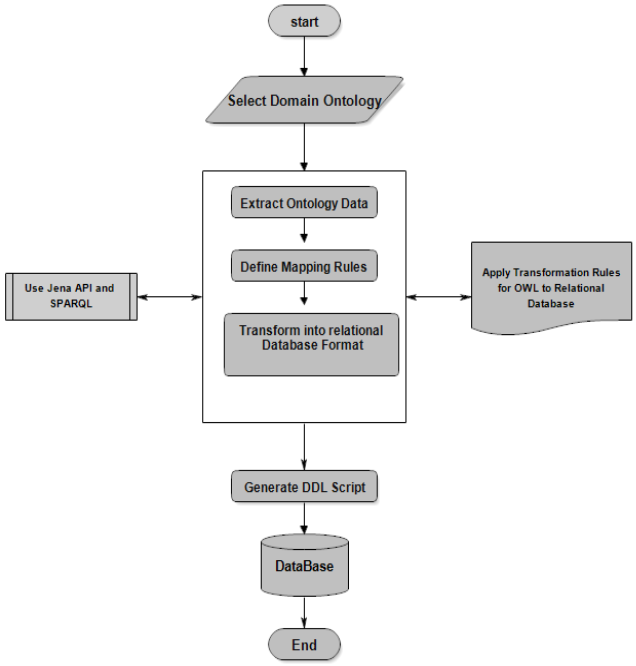


Fig. 1. System Architecture for Ontology Transformation to Relational Database (OWLMap)

B. Transformations Process for OWL to RDB

An algorithm has been developed to map ontology constructs into relational database format. Following that algorithm the transformation process is given below.

1) The given ontology or OWL file is first parsed to get Classes. Root class, super classes and subclasses are extracted.

2) Next Jena methods are applied to get two types of OWL properties i.e. object and data type. Data types of properties and restrictions are extracted as well.

3) Build database connection and transform this information into relational database format.

4) Classes and subclasses are transformed into separate tables and create one-to-one relationship among association classes according to mapping rules.

5) Then map properties as attributes of tables associated with corresponding class according to mapping rules.

6) Some properties are transformed into separate tables like multi-valued properties and properties having sub-properties.

7) Create separate metadata tables to store information about property restrictions.

8) Finally, ontology constructs are transformed into relational database.

C. Mapping Rules

Mapping rules are used to transform ontology to relational database. This section provides all the mapping rules used in OWLMap to transform ontology in to a database format. According to these rules, classes in ontology are transformed into relational database tables. The object type properties in selected ontology are transformed into columns or tables according to their relationship. Ontology data type properties are also transformed into columns or tables according to their values (single value or multi value).

1) OWL Classes:

Rule 1: Each OWL Class (Subclasses and association classes) will be transformed into a table in Relational database. Class name will become Table name. Table will be allocated a primary key. A table that relates to subclass is assigned a primary key a foreign key that reference to its "Super table" (one to one relationship between tables in relational database)

2) OWL Properties:

Rule 2: Single valued and functional object type property will be mapped into a foreign key in the table. The domain of the object property becomes the table. The range of the object property becomes another table. The name of the object property will be the name of foreign key linking two tables

Rule 3: Single valued and inverse of object type property will be mapped into a foreign key in the table that relates to range of object property and this key reference to primary key in the table that corresponds to the class specified as domain of object property. The name of the inverse object property will be the name of foreign key.

Rule 4: If Object type properties are multi-valued then they will be mapped into a separate table and will be assigned a primary key that's a combination of two foreign keys. One foreign key references the primary key of domain table and other to range table. The name of the object property will be the name of Table.

Rule 5: If object type properties are further divided into sub properties then they will be transformed into a table and their sub properties will be mapped into columns of that table. The name of the super property will be the name of table by adding prefix Prop_.

Rule 6: Single–valued data type property will be mapped into a column in the table that relates to the domain of data type property. The data type property's name will become the name of the column.

Rule 7: Multi–valued data type property will be mapped into a table and will be assigned a primary key that is a combination of corresponding column and the foreign key that reference to the domain table of data type property. The data type property's name will become the name of the Table

Rule 8: If data type properties are further divided into sub properties then they will be transformed into a table and their sub properties will be mapped into columns of that table. The name of the super property will be the name of table by adding prefix Prop_.

Rule 9: Data Type Conversion of Data type properties: We have converted data types of data type property from XSD to SQL, because OWL uses XSD data types. TABLE 1, shows how to convert different data types from XSD to SQL.

TABLE I. CONVERSION OF DATA TYPES TO DOL FROM XSD

XSD data Types	SQL Data Types
Short	SMALLINT
Integer	INTEGER
Negative Integer	INTEGER
Nonnegative Integer	INTEGER
Unsigned Int	INTEGER
Integer	INTEGER
Negative Integer	INTEGER
Nonnegative Integer	INTEGER
Unsigned Int	INTEGER
Long	INTEGER
Unsigned Long	INTEGER
Decimal	DECIMAL
Float	FLOAT
Double	DOUBLE PRECISION
String	CHARACTER VARYING
Normalized String	CHARACTER VARYING
Token	CHARACTER VARYING
Language	CHARACTER VARYING
NMTOKEN	CHARACTER VARYING
Name	CHARACTER VARYING
NC Name	CHARACTER VARYING
Time	TIME
Date	DATE
Datetime	TIMESTAMP
gYearMonth	DATE
gMonthDay	DATE
gDay	DATE
gMonth	DATE
Boolean	BIT
HexBinary	CHARACTER VARYING
AnyURI	CHARACTER VARYING

3) OWL Restrictions:

To preserve all information about ontological constraints, this information is stored in Meta data tables. Every type of restriction has its own table.

Rule 10: Some values from restriction maps to table having columns, restriction class (this column points to the table of the related restriction resource class), property (includes the property concerned), domain class and range class of property. All values from restriction maps to table having columns, restriction class (this column points to the

table of the related restriction resource class), property (includes the property concerned), domain class and range class of property. Has value restriction maps to table having columns, restriction class (this column points to the table of the related restriction resource class), oN property (includes the property concerned), domain class and range class of property. In case of "Has value Restriction" Meta data table, a column "value" is added for storing the value of restricted resource of related property.

Rule 11: Inverse functional property will be mapped to unique constraint on the corresponding column. And required Property will be mapped on the corresponding column as Not Null Constraint.

IV. EXPERIMENT

Number of experiments has been performed for ontology to relational database transformation using OWLMap. Different ontologies from multiple domains are presented to our OWLMap tool for transformation. Experiments show that proposed approach is fully automatic, effective and quick. This approach is lossless as well and performs the transformation successfully.

Fig. 2. Class hierarchy available in pizza ontology

In this section, we have presented automatic transformation process taking famous "Pizza Ontology" as an input. Pizza ontology is downloaded from standard Website of Stanford University. Pizza ontology is developed at Manchester. It has often been considered as important ontology for learning basic concepts of ontology and OWL language [18] and [19]. This ontology has been chosen because pizzas are widely understood in all cultures or across the world. The pizza ontology includes most of OWL features. These ontology concepts are used to present main components of pizza domain, as illustrated in fig 2. This ontology has number of Object type properties, data type properties and their sub properties, shown in fig 3 and 4.

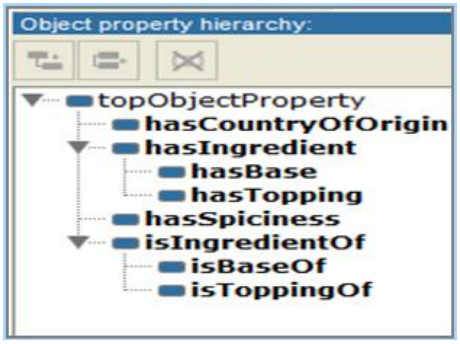


Fig. 3. Object type properties in pizza ontology

First, download the pizza ontology, explore it with the help of Protégé and check its consistency by using reasonor. Next, with the help of Jena API, extract all the information about ontological constructs e.g. classes, sub classes, object properties, data type properties, their domain and range, data types, restrictions etc. according to mapping algorithm.

Fig. 4. Data type properties in pizza ontology

In the next step, choose SQL Server 2008 R2 to store this information in relational database format. Database is created and connection with the database is established. Then implement defined mapping rules to transform ontology into relational database format. According to rule 1, transform all ontology classes into tables in RDB and class name become table name, assign a primary key as shown in fig 5. A table that relates to subclass is assigned a primary key and a foreign key that reference to its "Super table" (one to one relationship between tables in relational database) as shown in fig 6.

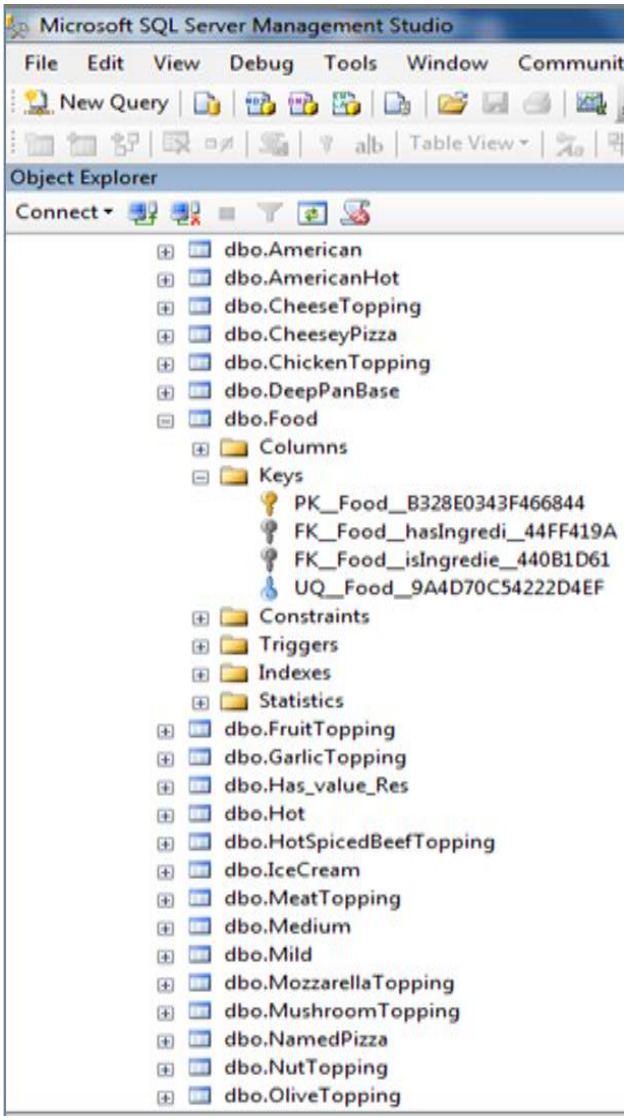


Fig. 5. Conversion of ontology classes into tables in RDB

Fig 6, explains the one to one relationship between class and its related subclass. In Pizza ontology "Pizza class" is constructed as subclass of "food class". So when Pizza class is transformed into relational database format, a Pizza table is created with one to one relation to its super class table called Food. This mapping rule is applied in all association classes while transforming into relational database format.

After mapping classes and subclasses into relational database format, transformed object type properties according to defined mapping rules. According to mapping rules, transform single valued and functional object type property into foreign key in the table that relates to domain of object type property and this key reference to primary key in the table that relates to the range class of object type property as illustrated in fig 7 and 8.

In fig 7, "Has Spiciness" is functional property and has specified "Spiciness" class as domain and range of this property. The object type property transforms to a foreign key in the table. That table represents a class specified as the domain of the object property. In fig 8, the key that reference the primary key in the table is related to the class specified as the range of object property. We have transformed object type property that is "inverse functional" and single valued into a foreign key.

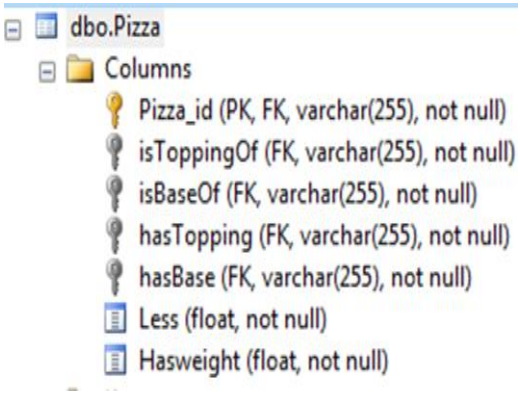


Fig. 6. Conversion of ontology subclass into table in RDB

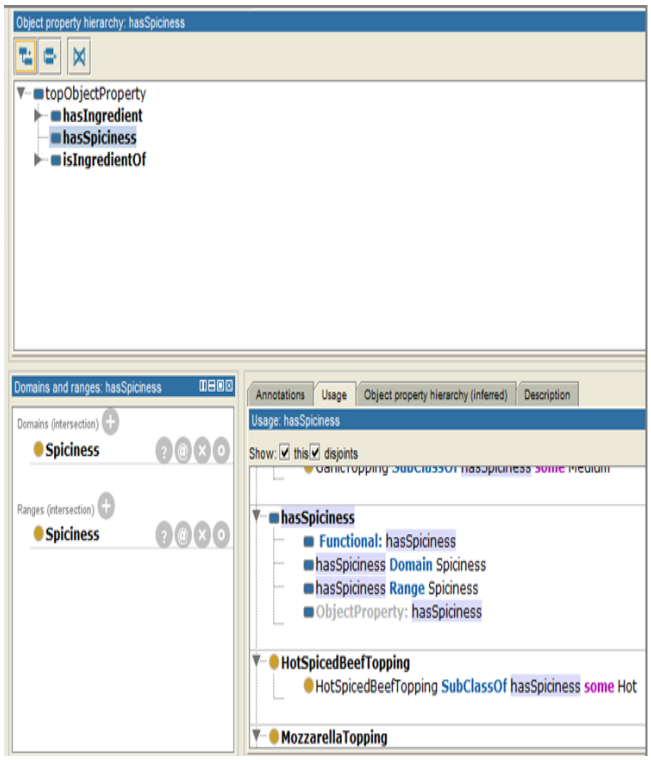


Fig. 7. Functional object type property in pizza ontology

The table that relates to range class of object type property and this key reference to primary key in the table that relates to domain of object type property as illustrated in fig 9 and 10.

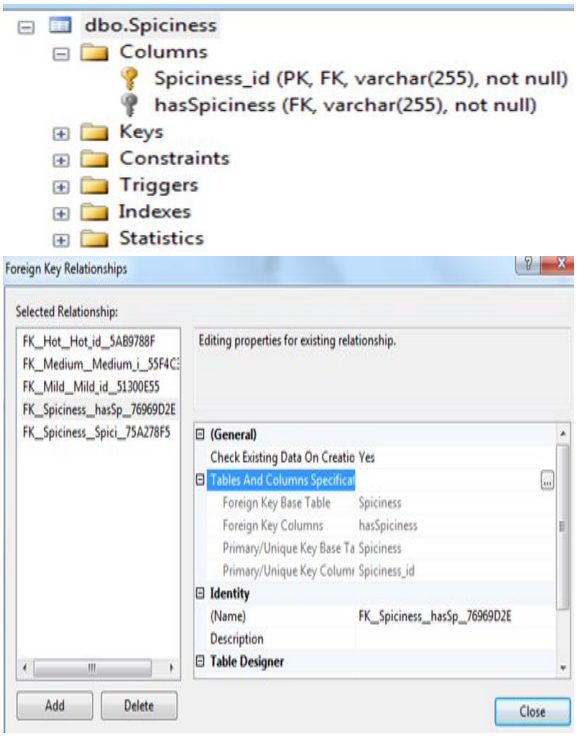


Fig. 8. Transformation of single valued and functional object property

In fig 9, "is Base of" object type property is inverse of "has Base" object type property. If object property has sub properties e.g. "has Ingredient" has two sub properties "has Base" and "has Topping' then super property maps to a table in relational database and its sub properties maps to a column in corresponding table as shown in fig 11.

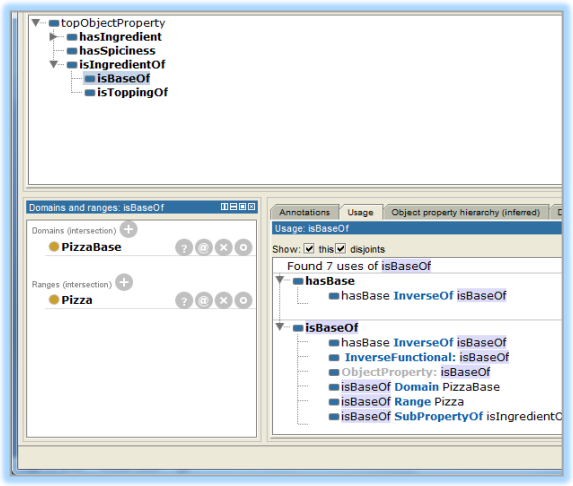
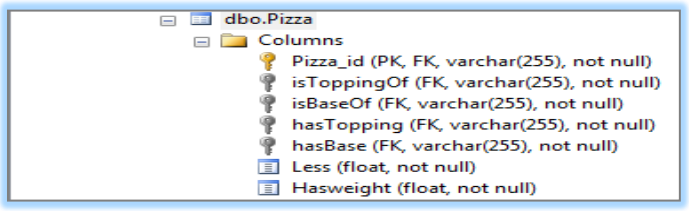


Fig. 9. Inverse functional and single valued object property



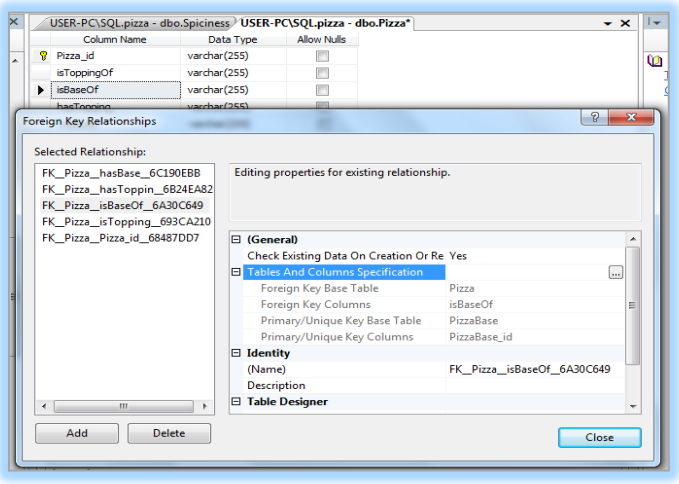


Fig. 10. Conversion of object function property

After mapping object type properties, OWLMaptransfors data type properties and their data types into relational data base format. If data type property is single valued e.g. In Pizza ontology "has Price" is a data type property and has specified "Pizza" class as its domain and has data type "float". In fig 12, the property in ontology maps to a column in the table that relates to the class specified as domain of data type property. In table 1, the column data type which is specified as range of data type property converted from XSD data type to SQL data type. If data type property has sub properties e.g. in pizza ontology "has Weight" data type property has sub property "Less", so this property maps to a table in relational database and its sub properties maps to a column in related table as shown in fig 13. While converting OWL ontology into relational database format, we want to preserve all information of ontological constraints. For this purpose we have saved this information in special Meta data tables. Each type of restriction has its own table as explained above in mapping rules.

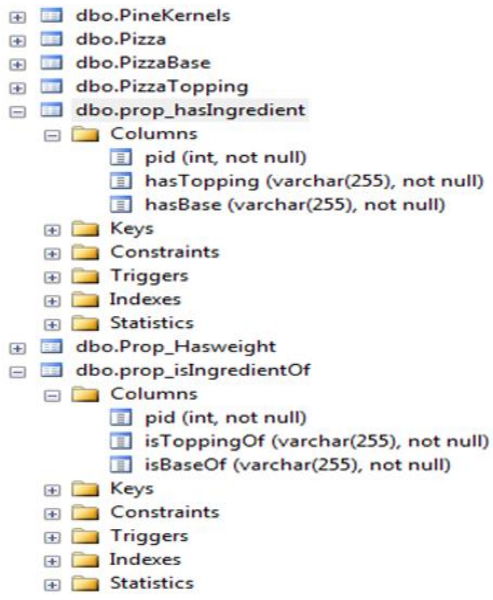


Fig. 11. Conversion of object type property that is further divided into sub properties

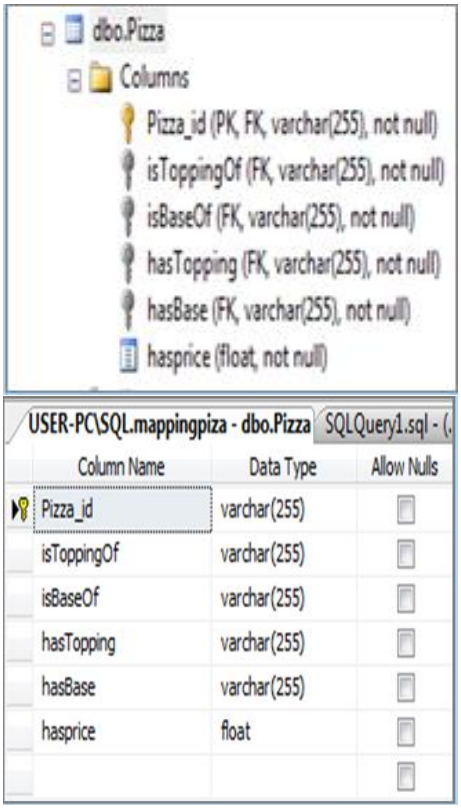


Fig. 12. Conversion of single valued data type property and its data type

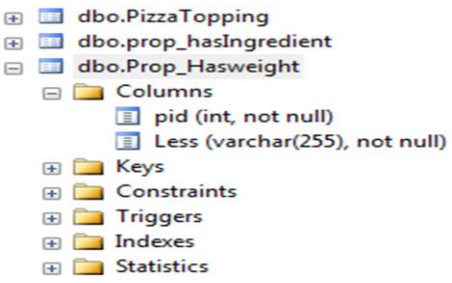


Fig. 13. Conversion of data type property that is further divided into sub properties

Results | Messages

	Some_id	ResCLass	Prop	Propdomain	PropRange
1	1	Pizza	hasBase	Pizza	PizzaBase
2	2	OliveTopping	hasSpiciness	Spiciness	Spiciness
3	3	GarlicTopping	hasSpiciness	Spiciness	Spiciness
4	4	NutTopping	hasSpiciness	Spiciness	Spiciness
5	5	MushroomTopping	hasSpiciness	Spiciness	Spiciness
6	6	ChickenTopping	hasSpiciness	Spiciness	Spiciness
7	7	AmericanHot	hasTopping	Pizza	PizzaTopping
8	8	MozzarellaTopping	hasSpiciness	Spiciness	Spiciness
9	9	IceCream	hasTopping	Pizza	PizzaTopping
10	10	American	hasTopping	Pizza	PizzaTopping
11	11	HotSpicedBeefTopping	hasSpiciness	Spiciness	Spiciness
12	12	SultanaTopping	hasSpiciness	Spiciness	Spiciness

Fig. 14. Meta data table "some values from restrictions" after conversion

Fig 14, shows some values from the restrictions are mapped to table having columns restriction class (this column points to the table of the related restriction resource class), property (includes the property concerned), domain and rRange. In case of "Has value Restriction" when we have created Meta data table, a column "value" is added for storing the value of restricted resource of related property.

V. TESTING

To test the performance of OWLMap tool, ten different ontologies have been taken from standard web site of Stanford University. These ontologies have different sizes, and from different domains. Testing of developed tool with different types of ontologies increases its efficiency and reliability. In table 1; some important specifications of our machines are given. These components play vital role in process of testing.

TABLE II. MACHINE SPECIFICATIONS

Processor	**CPU speed**	**OS**	**Memory**	**System type**
Intel core i5	2.40GHz	Windows 7	4 GB	64 bit operating system

We have also observed the time required for conversion from ontology to database by developed tool. It is observed that conversion time is different for large and small ontologies. For large ontologies, it takes about 25 seconds and

for small ontologies it takes only 10 seconds in transformation. In table 2, detailed information about sample ontologies and their components are given. This information plays an important role while testing of given approach. It

becomes very easy to check the efficiency of developed tool by comparing this information.

TABLE III. DETAIL ABOUT SAMPLE ONTOLOGIES

Sample Ontology	Description	Number Of Classes	Object Type Properties	Data Type Properties	Restrictions
Pizza	About pizza and its components	>80	7	3	Yes
University	About university domain	29	6	2	Yes
Camera	About camera and its components	11	7	8	Yes
Trade	Define trade system	>70	20	3	No
Travel	Defines tourism	33	6	4	Yes
Event	About event handling	>60	7	10	No
Delegation	About management	18	5	3	Yes
Education	Defines education domain	36	12	>30	No
Car Advertising	About car advertising	10	3	8	No
Wine	About wine components	>15	3	1	Yes
Db1	About Library system	7	3	14	No
Docdb	About hospital	4	4	>10	No

TABLE IV. EVALUATION OF SUGGESTED TOOL WITH THE HELP OF SAMPLE ONTOLOGY

Converted ontology concepts	Classes	Subclasses and their relationship	Object properties	Data type properties and their data types	Restrictions
Pizza Ontology	94%	Yes, 40%	98%	Yes, all	Yes, converted to Metadata tables
Edu Ontology	100%	Yes, 60%	100%	Yes, all	Yes, converted to Metadata tables
Trade Ontology	100%	Yes, 60%	98%	Yes, all	Null
Travel Ontology	98%	Yes, 60%	94%	Yes, all	Yes, converted to Metadata tables
Event Ontology	100%	Yes, 60%	60%	60%	Null
Delegation Ontology	96%	Yes, 60%	100%	Yes, all	Yes, converted to Metadata tables
Education Ontology	100%	Yes, 60%	20%	70%	Null
Car Advertising Ontology	100%	Null	100%	Yes, all	Null
Wine Ontology	98%	Yes, 40%	100%	Yes, all	Yes, converted to Metadata tables
Camera Ontology	100%	Yes, 80%	100%	Yes, all	Yes, converted to Metadata tables

VI. Conclusion and Future Work

As the semantic Web is gaining importance, there is a need of an efficient approach to map all ontology information into relational database so that it can be queried easily. Therefore, we have developed a tool OWLMap for automatic and lossless transformation of ontology into relational database. This transformation approach can map all the constructs of ontology including "sub properties" that are not handled before in any transformation approach. For lossless transformation of ontology constructs into relational format, mapping algorithm and rules are defined. According to these mapping rules, ontology classes should be transformed into tables, object type properties should be mapped into columns or tables, data type properties should be mapped into columns or tables according to mapping rules and restrictions must be stored into meta data tables.

OWLMap can help in mapping ontological data into relational databases that can be further utilized in different applications. It is easy to access heterogeneous and distributed information. This approach will play an important role in Advance Querying / Query Optimization. Other benefits of this approach are quick retrieval of data and schema Integration. This approach is capable to automatically transform most of ontology constructs into relational structure. In future, it is required to transform some other ontological information i.e. Class complements, comments, and enumerated or intersection classes. Reflexive and Irreflexive properties of OWL also need some attention in future.

References

[1] V. Jain, and S. V. A. V. Prasad, "Mapping Between RDBMS And Ontology: A Review", International journal of scientific & technology research volume 3, issue 11, 2014.

[2] D. Moldovan, M. Antal, D. Valea, C. Pop, T. Cioara, I. Anghel, and I. Salomie, "Tools for Mapping Ontologies to Relational Databases: A Comparative Evaluation", In Intelligent Computer Communication and Processing (ICCP), 2015 IEEE International Conference, pp. 77-83.

[3] I. Astrova, and A. Kalja, "Storing OWL Ontologies in SQL3 Object-Relational Databases", In AIC'08: Proceedings of the 8th conference on Applied informatics and communications, pp. 99-103.

[4] H. Afzal, T. Naz, and A. Sadiq, "A Survey on Automatic Mapping of Ontology to Relational Database Schema", Research Journal of Recent Sciences, Vol. 4(4), 66-70, 2015.

[5] S. H. Tirmizi, J. Sequeda, and D. Miranker, "Translating SQL Applications to the Semantic Web", In Proceedings of the 19th International Conference on Database and Expert System Application: 450-464, 2008.

[6] Z. Wu, G. Eadon, S. Das, E. I. Chong, V. Kolovski, M. Annamalai, and J. Srinivasan, "Implementing an Inference Engine for RDFS/OWL Constructs and User-Defined Rules in Oracle, In Proceedings of IEEE 24th international Conference on Data Engineering: pp. 1239-1248. Mexico, Cancum, 2008.

[7] C. D. Barranco, J. R. Campana, J. M. Medina, and O. Pons, "On storing ontologies including fuzzy datatypes in relational databases", In Proceedings IEEE International Conference on Fuzzy System, July 23-26, 2007: 1-6.

[8] R. Goodwin, and J. Y. Lee, "Ontology Management for Large Scale Enterprise". Electronic Commerce Research and Applications5 (1): 2-15, 2006.

[9] N. Zina, and N. Kaouther, "Automatically building database from biomedical ontology"

[10] I. Astrova, N. Korda, and A. Kalja, "Storing OWL Ontologies to SQL Relational Databases", International Journal of Electrical, Computer, and Systems Engineering 1(4), 2007.

[11] A. Gali, C. X. Chen, K. T. Claypool, and R. Uceda-Sosa, "From ontology to relational databases", In Proceedings of International Workshop on Conceptual-Model Driven Web Information Integration and Mining: 278-289. Shanghai, China, 2005.

[12] E. Vysniauskas, and L. Nemuraite, "Transforming Ontology Representation from OWL to Relational Database", Information Technology and Control, 2006, 35(3A): 333–343.

[13] E. Vysniauskas, and L. Nemuraite, "Mapping Of OWL Ontology Concepts To RDB Schemas" In proceedings of the 15th International Conference on Information and Software Technologies Kaunas, Lithuania, April 23-24, 2009: 317-327.

[14] W. Hu, and Y. Qu, "Discovering Simple Mappings between Relational Database Schemas and Ontologies", In Proceedings of the 6th international the semantic Web and 2nd Asian conference on Asian semantic Web conference: 225-238, 2007.

[15] D. D. B. Saccol, T. D .C. Andrade, and E. K. Piveta, "Mapping OWL Ontologies to Relational schemas", In proceeding of IEEE International Conference on Information Reuse and Integration (IRI), 2011.

[16] E. Vysniauskas, L. Nemuraite, R. Butleris and B. Paradauskas, "Reversible Lossless Transformation from OWL 2 Ontologies into Relational Database", Information Technology and Control (4), 2011.

[17] G. Antoniou, P. Groth, F. V, Harmelen, and R. Hoekstra, "Semantic Web Primer" , MIT press/London, 2012.

[18] M. Horridge, N. Drummond, J. Goodwin, A. Rector, R. Stevens, and H. H. Wang, "The Manchester OWL Syntax"

[19] R. Sivakumar, and P.V. Arivoli, "Ontology Visualization Protégé Tools-A Review", International Journal of Advanced Information Technology 1, no. 4 , 2011.

Vismarkmap – A Web Search Visualization Technique through Visual Bookmarking Approach with Mind Map Method

Abdullah Al-Mamun
Department of Computer Science & Engineering
Daffodil International University
Dhaka, Bangladesh

Sheak Rashed Haider Noori
Department of Computer Science & Engineering
Daffodil International University
Dhaka, Bangladesh

***Abstract*—Due to the massive growth of information over the Internet, Bookmarking becomes the most popular technique to keep track of the websites with the expectation of finding out the previously searched websites easily whenever are needed. However, present browser bookmark systems or different online social bookmarking websites actually do not let the users to manage their desired searches with appropriate method, so that the users could easily recognize or recall the previously searched websites and its content with the bookmark whenever they are in need. In this paper, a new approach of bookmarking technique has been proposed which will let the users to organize their bookmarks with some features using mind map, a scientifically approved mental model that will help the users to recall easily the previously searched websites' information from the bookmarks and will minimize the tendency to revisit or research the website using the search engines. Basically, the proposed system is more than a mind map as it provides more flexibility to organize the bookmarks.**

Keywords—hci; visual bookmark; information retrieval; mind map; visualization

I. INTRODUCTION

Since the World Wide Web releases to the public, Internet continues to grow and becomes a very important source of information. For seeking information on the web users follow several different strategies such as direct navigation, navigation within a directory and navigation using a search engine; among which the search engines make the information searching process more easier [10] than others.

However, in the user perspective, a very crucial problem is being provided very less care, which is about coming back to information they have previously found [5]. During a particular search session a search engine might find the desired search results; though, finding the same result users might require to search or navigate again with extensive effort, since due to the rapid increment of the registered domain (approximately 350 million) since last decade [13], the World Wide Web is becoming massive and largely disorganized. Therefore, finding out a previously visited webpage might lead to an unsatisfying and unproductive experience. In addition, different search engines provide the search results of different web sites in list based approach. However, users prefer to devote very small amount of time for the navigation inside web sites, rather than using the search results list they prefer to jump from one site to another [10]. As the quality of information on the Web varies and the user has to make a judgment within very short period of time before jumping from one site to another, users must need to have a quick visualization model in which they will be able to compare among the similar or dissimilar search results to find out their expected web site very quickly, which might provide them a satisfying information searching experience as they will need to put less mental effort.

On the other hand, to facilitate the process of finding previously searched results, present commercial web browsers such as Google Chrome, Mozilla Firefox and Microsoft Internet Explorer are providing almost identical functions for returning to these pages, such as Back, the history list, and bookmarks. The aforementioned mechanisms should greatly be used, as according to the past research [11] a person browses mostly (60% of all the pages) previously visited pages. However, due to several problems with these mechanisms, bookmark and history systems are rarely used [11, 12]. One of the main reasons is the scattered and un-integrated method of the browsers re-visitation systems [5]. The major functions such as Back, history and bookmarks, all use different models, user interfaces, and provides numerous techniques of organizing and visualizing groups of candidate pages. As a result, recalling a website from the history list becomes difficult.

To overcome the aforementioned problems, in this paper, a new bookmarking technique has been proposed which will let the users to organize their bookmarks using mind map, which is a scientifically approved mental model. With the help of the mind maps, it will help to find back easily their previously searched websites by minimizing their (users) dependency on the search engines, which will provide a better searching experience.

The rest of the paper is organized as follows. Some of the important related researches are reported in section 2. In section 3, the solution against the traditional bookmarking systems is proposed. The proposed solution is explained in details in section 4. In Section 5 and 6, the evaluation of the experiment and the related discussion is presented respectively.

II. Related Research

At present, for the bookmarking process user use either renowned web browsers' default bookmark system or different social bookmarking websites, such as delicious, Pinterest and Diigo etc. In present commercial web browsers such as Google Chrome, Mozilla Firefox and Microsoft Internet Explorer, we bookmark our favorite websites with tags. The Delicious allows users to bookmark the useful searches with tags and to share with people. Unfortunately from the researches [3] it has been found there is a natural convergence of the tags on the Delicious website. Even so, these tags lack embedded semantics, so when observing Delicious as tagging platform, the common ambiguity problems of tagging (such as synonyms or false tags) [4] are encountered. In Diigo[1] users can create bookmark list of web sites along with tags and also are able to keep notes with specific fragments of a webpage which are visible when those websites are revisited. With the Pinterest users are able to bookmark the searches along with thumbnail preview, tags and personal notes. Moreover, in past, several applications were implemented to organize and visualize the web search results. Shaun Kaasten and Saul Greenberg [5] proposed an integrated system which worked in Microsoft Internet Explorer. Basically they designed a new view of history list to represent the visited pages with visual thumbnails along with titles and respective URLs. It includes two types of bookmarking process such as *Implicit* and *Explicit* bookmarking with the help of *Dog Ear* metaphor. The most visited pages are marked automatically with *Implicit* bookmark and *Explicit* bookmarking is done by the users. In addition, it supports searching the bookmarks based upon the most visiting frequency. David Gotz proposed a system [9] where user can drag and drop URL of a link from browser to bookmark and can create a hierarchical tree of connected nodes of URLs. Besides, Users can store full webpage or a particular segment of a webpage as a node in the tree.

After analyzing the aforementioned works, it can be concluded that most of the researches followed either simple data clustering or symbolic representation or spatial representation with the help of tree to organize the bookmarks. However, these approaches are not appropriate enough to visualize information and knowledge of the bookmarks in coherent approach as those are not following the appropriate cognitive theory.

III. Proposed Solution

As user has to decide very quickly before jumping from one bookmarked site to another by comparing among the similar or dissimilar bookmarks very quickly, in the proposed solution the focus is put on the visualization of the information and knowledge of the bookmarked websites following the cognitive fit theory [1]. Proper visualization mitigates the limitation of the working memory and helps the learner overcome problems during the process of learning and problem solving. According to Sweller and Chandler [14] visualizations may reduce cognitive load and in [15], [16], and [17], it is mentioned that visualization expand the capability of an individual's memory for coping with complex cognitive task requirements. Cox [15] and Scaife & Rogers [18] mentioned another significant reason; visualizations can enhance our processing ability by visualizing abstract relationships between visualized elements and may serve as a basis for externalized cognition.

To visualize information and knowledge together, in the proposed solution *Synergistic* approach is followed. The *Synergistic* approach aims at integrating knowledge and information visualization in coherent approach [8]. It is claimed that map based approach such as concept map can be used for mapping and managing conceptual knowledge among information [6]. According to cognitive fit theory [1], it is also known that, graphs work as representations of the spatial problems because they present spatially related information which emphasizes relationships in the data. Even though, concept map is helpful for organizing knowledge, it is not considered as complete mental model as it requires imagery-based elements to comprehensively represent the knowledge [2]. In addition, research [5] shows organizing bookmarks with thumbnail images, makes the process of scanning the bookmark list easy for the users. Therefore, in this paper a new bookmarking technique *VisMarkMap* is proposed with the help of mind map as it is very effective way of the *Synergistic* approach to represent information and knowledge coherently and removes the shortcomings of the concept maps; because, with the help of mind map [Figure 3] one can represent the similarity or dissimilarity among the bookmarks along with visual imagery elements and tags. In addition, the proposed technique also supports the visualization of the quality of the bookmarks based upon the most visiting frequency and the five star rating, provided by the users though ranking mechanism. On the other hand, discrete values cannot be presented directly with graphs such as mind maps or concept maps [1]. For symbolic problem representations one can use tables or list based approach because they present symbolic information and emphasize discrete data values. Therefore, for visualizing all mind maps' bookmarks together, a tag list based approach [Figure 8] has been chosen to give the users an overall knowledge about all the bookmarks of all mind maps.

[1] https://www.diigo.com/

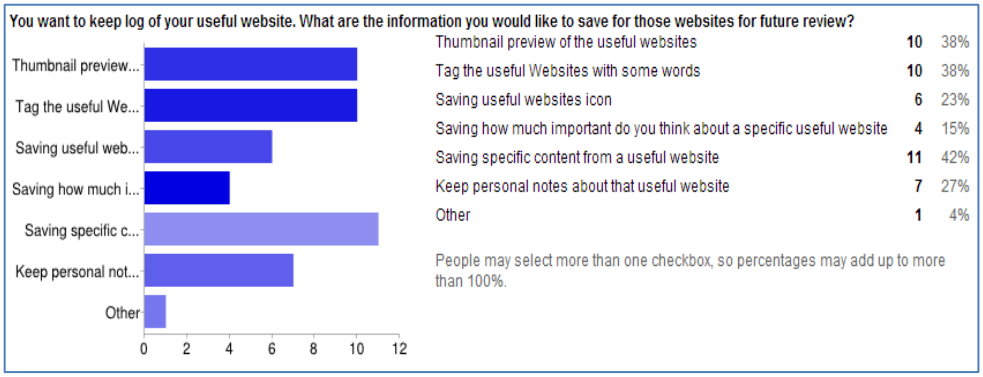


Fig. 1. Initial survey result for bookmark components

IV. VisMarkMap-Visual Bookmarking with Mind Map Method

At the beginning a survey has been conducted in order to find out the expected information to be visualized about a website bookmark, in which 26 people (students and teachers of different college and universities) participated. From the outcome of the survey [Figure 1] it is found that most users expect to see the thumbnail, the associated tag and specific content kept from that respective web site. Besides, users also wanted to keep personal notes about the bookmarked web sites. Therefore, the proposed visual bookmarking system (VisMarkMap), which is implemented as a browser extension for Chrome, allows users to organize their bookmarks along with mind maps with above mentioned features. The overall process of the VisMarkMap is depicted in Figure 2.

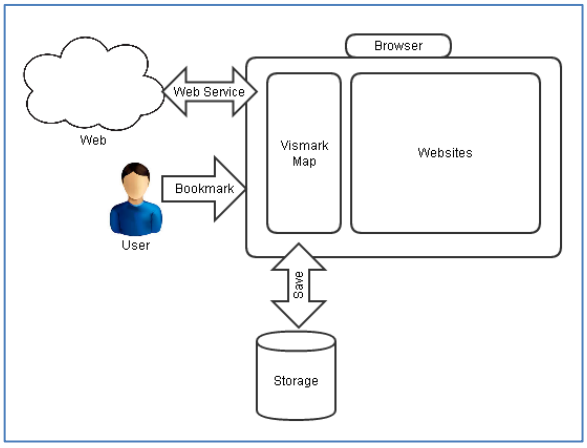


Fig. 2. VisMarkMap architecture

During a search session, user can interact with the visual bookmark button on the browser and the bookmark panel will then be shown on the left side of the browser [Figure 9]. Then a user can start creating new mind map with *NewMap* button and create new node with *NewNode* button. After that, the user will select the desired URL (website or image or video) from the browser address bar and will perform a simple drag & drop operation to fetch the URL on the bookmark panel [Figure 3]. The VisMarkMap will then utilize a web service to create the necessary thumbnail and big preview image or small preview clip for video bookmark associated with the bookmarked URL.

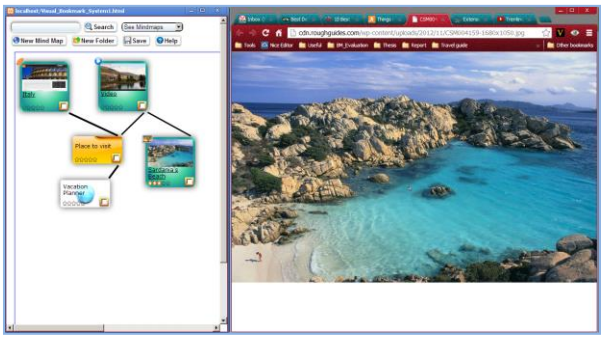


Fig. 3. Bookmarking in mind map

In addition, Figure 4, Figure 5 and Figure 6 show a user can select the desired text from bookmarked website and can keep as a note along with the big preview image, which mitigates the necessity of visiting the respective website again. This system will also assist users to keep date and to perform necessary tagging on the bookmarks as well as to create groups by allowing making desired relationships among the similar bookmarks to form a mind map.

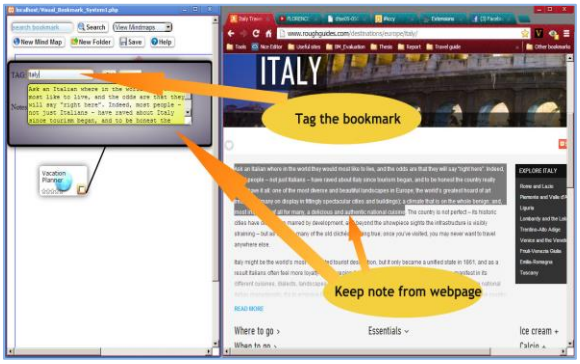


Fig. 4. Tagging and keeping notes

There is also a five star rating mechanism with which users can mark the importance rate of the bookmarks in a mind map [Figure 7]. Besides, based on the most visiting frequency the bookmarks will be highlighted on the mind map [Figure 7]. The summary of the mind maps will also be shown as tag

boxes on the home panel [Figure 8], which contain the user defined tags of the bookmarked sites.

Fig. 5. Preview of a bookmarked website

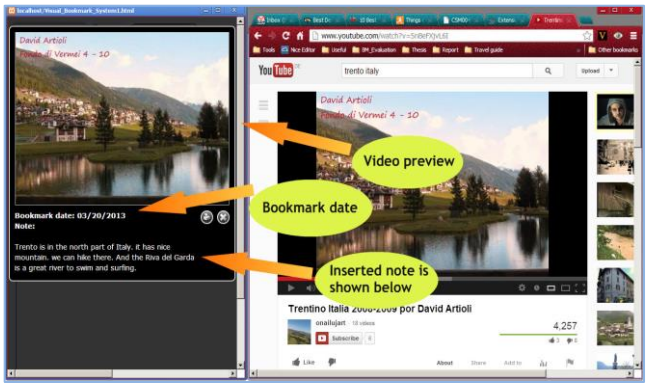


Fig. 6. Preview of a bookmarked video

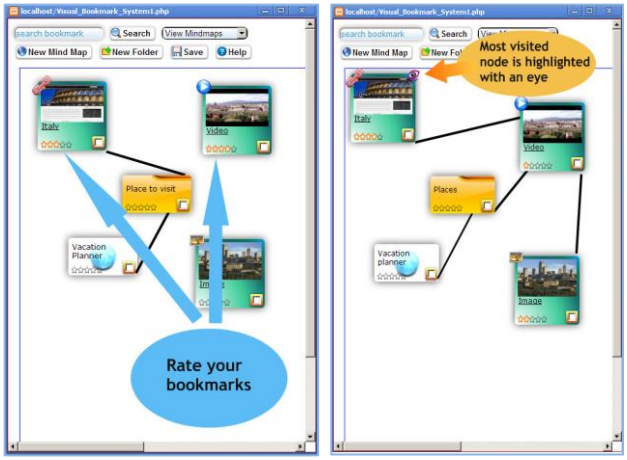


Fig. 7. Five star rating and highligting most visited bookmarks

Afterwards, when user in need of looking back into previously searched information, they can search within this new system using keyword just like searching using web search engines, and then the system will respond with boxes filled up with tags that matched with respective keywords, where each box represents a mind map. User can browse and edit the mind maps by clicking on the respective tag boxes [Figure 8].

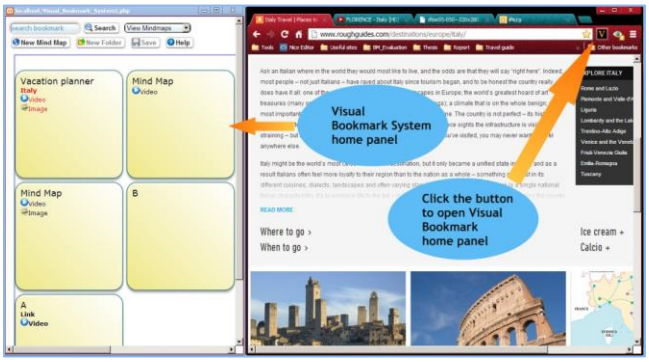


Fig. 8. All mind maps with bookmark tags

Finally, the features of the proposed system help users to organize their bookmarks in such a way that users are able to find back easily their previously searched websites' most expected information within their browser that will minimize the probabilities of searching them again with the help of search engines or re-visiting them.

V. EVALUATION

In order to test whether people have difficulty with the concept of proposed system, a short preliminary study has been conducted in which 15 people were requested to use the first prototype of the VisMarkMap as a bookmarking system. Here, the people who were interviewed are students and teachers of school and universities. The experiment is conducted in two phases. In the first phase, initially, users were introduced briefly about the VisMarkMap System. Then users were given a specific scenario [Figure 9] to perform some searching and bookmarking both with the traditional bookmark systems (i.e. chrome browser bookmark, Delicious, Pinterest) as well as with this new visual bookmark system (VisMarkMap).

The initial goal of the first phase was to measure the performance of the users on using the user interface of the first prototype of the new visual bookmark system. In this phase the usability of the user interfaces of the new system is also verified. The test is designed using the format used for formative evaluation by [19]. For measuring the performance, some significant quantifiable usability goals have been chosen [Table 1]. While performing the scenario using the new visual bookmark System, users were directly observed to check how many attempts they exactly need to perform a specific task.

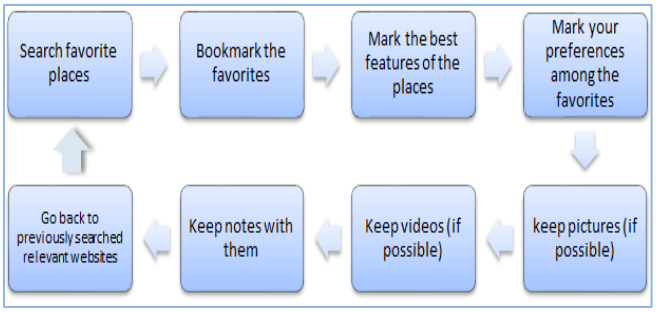


Fig. 9. Vacation Planner Scenario

TABLE I. USABILITY GOALS TO MEASURE

Usability Attribute	Measuring Instrument	Value to be Measured	Worst Acceptable Level	Planned Target Level	Best Possible Level
Initial use	Create a new mind map	Number of attempt	3	2	1
Initial use	Create a custom node	Number of attempt	3	2	1
Initial use	Create a bookmark	Number of attempt	3	2	1
Initial use	Connect the nodes in a mind map	Number of attempt	3	2	1
Initial use	Tagging the nodes	Number of attempt	3	2	1
Initial use	Preview the nodes	Number of attempt	3	2	1
Initial use	Delete a connection between nodes	Number of attempt	3	2	1

In the second phase, users were requested to perform the same scenario again after couple of days. During the second phase users are again observed to find out the tendency of using search engines to find out the previously searched websites. After the completion of the second phase, questionnaire method was used to measure the usability of the new system against the all other well-known bookmarking systems such as traditional browser bookmark, Delicious and Pinterest. The questionnaire was designed carefully using several standard software usability measuring questionnaire [20, 21, 22], to measure the four significant areas of the usability such as,

- Usefulness and effectiveness
- Ease of use
- Ease of learning
- Satisfaction

The questionnaire was built based upon the Likert scale and the respondents were allowed to indicate their agreement or disagreement with a 7 point scale.

A. Evaluation Results

a) First Phase

In Figure 10 to Figure 16, the outcome of the user observation during the first phase is presented. It is found, in most cases majority of users were able to finish the various tasks on their first trial, without any mistake, whereas, in some cases, one third of the users made at most one mistake and were able to finish the tasks at most with two trials.

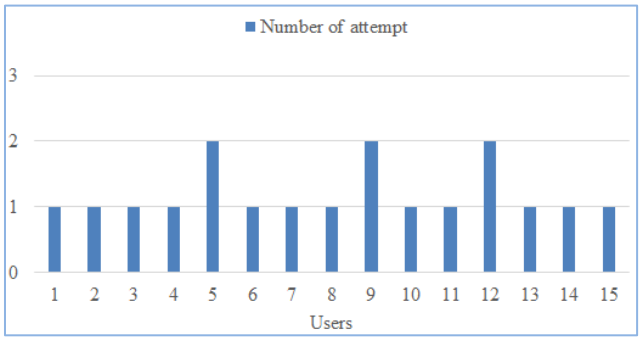


Fig. 10. Create a new mind map

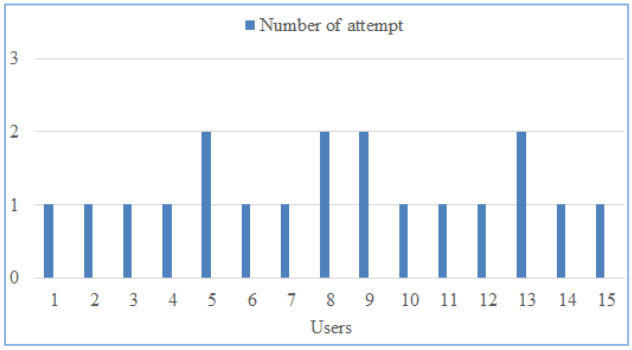


Fig. 11. Create a custom node

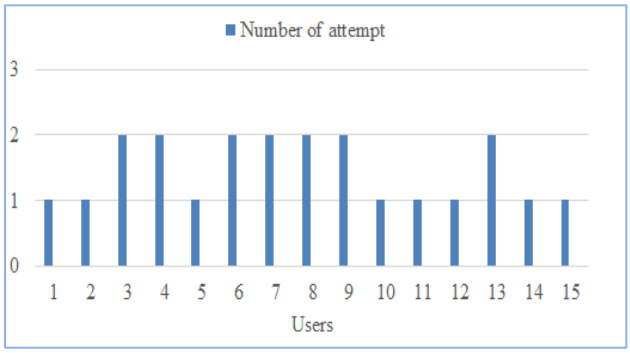


Fig. 12. Create a bookmark

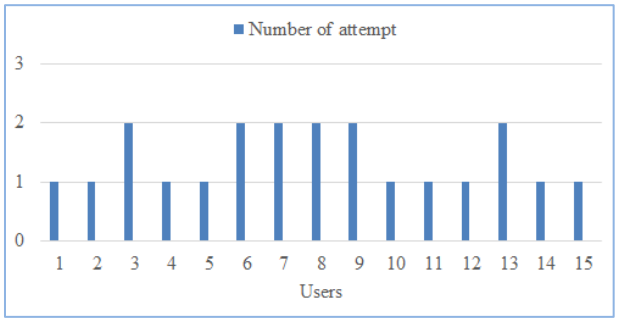


Fig. 13. Connect the bookmark nodes

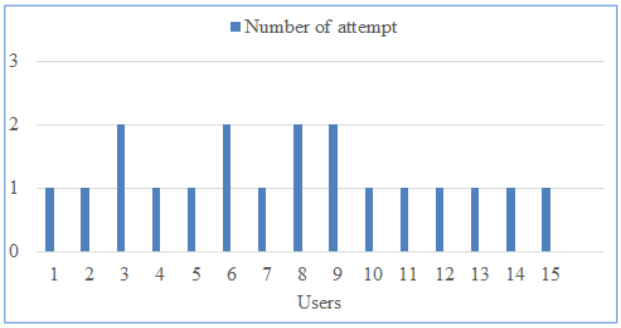


Fig. 14. Tagging the nodes

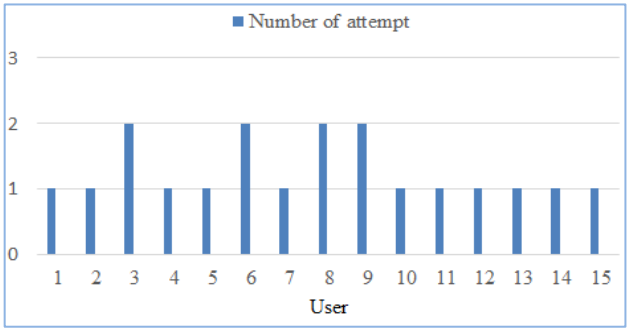


Fig. 15. Preview the bookmarks

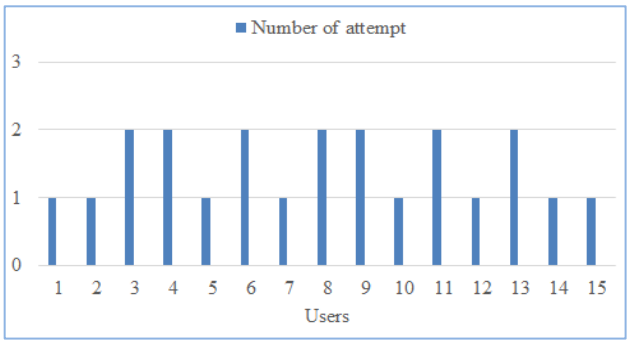


Fig. 16. Removing a connection between nodes

b) Second phase

During the second phase of user observation it has been recorded how many times a particular user attempt for search engines while using the existing bookmark systems and the new visual bookmarking system, to find out previously searched websites' information. The experiments results for Delicious, Pinterest and new system are shown on Figure 17, Figure 18 and Figure 19 respectively.

Figure 19 shows more than 50% users (8 out of 15) did not attempt for search engines at all while to find out previously searched websites' information after understanding the features provided by the new system about which they came to know during the first phase, which helps users by minimizing their effort in case of comparing and finding the desired websites from search results returned as a list by the search engines. Using the new bookmark system, the percentage of people who attempted for search engines is reasonably lower compared to the other systems. It was also observed, most of the time almost all of the users attempted for search engines while they were using chrome browser bookmark system.

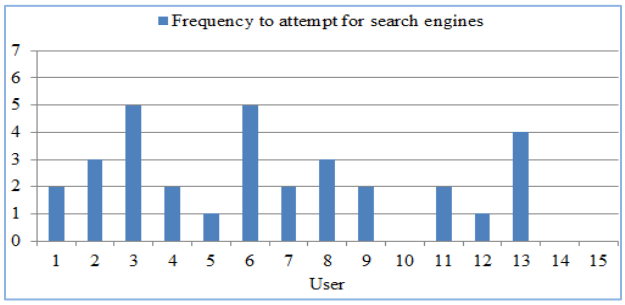


Fig. 17. Frequency to attempt for search engines using Delicious

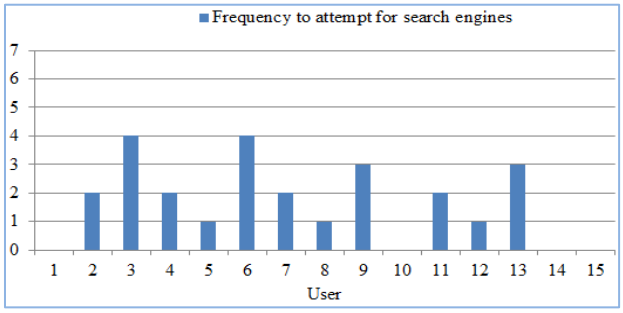


Fig. 18. Frequency to attempt for search engines using Pinterest

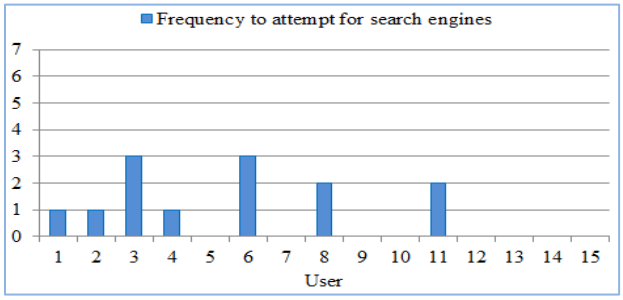


Fig. 19. Frequency to attempt for search engines using the new system

From the collected responses from questionnaire phase it has been found that most people agreed upon the fact that with the VisMarkMap they are able to bookmark according to their expectation (60% agreed & 20% strongly agreed) [Figure 20]. Significant numbers of people think VisMarkMap is making them more productive (30% strongly agreed & 40% agreed) [Figure 20] and most people think it as more useful (40% strongly agreed and 40% agreed) than the other bookmarking systems [Figure 20].

A key number of users found the VisMarkMap is easy to use (40% strongly agreed, 10% agreed & 20% agreed more than disagree) and easy to learn (50% strongly agreed & 40% agreed) [Figure 21]. Most importantly, majority of people (40% agreed & 30% strongly agreed) were satisfied about the overall functionalities of the first prototype of the new visual bookmarking system (VisMarkMap) [Figure 22].

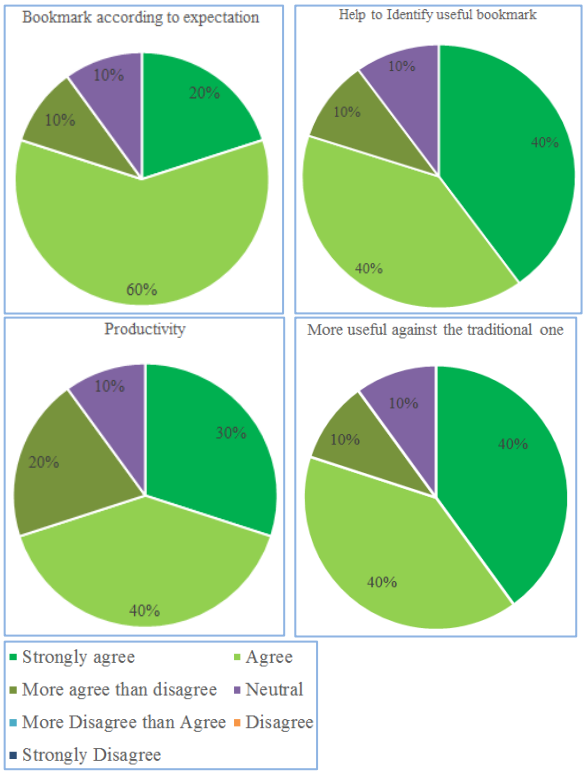


Fig. 20. Usefulness of VisMarkMap

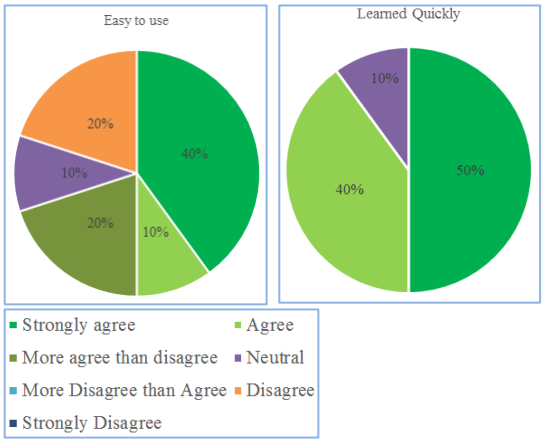


Fig. 21. Ease of use of the VisMarkMap

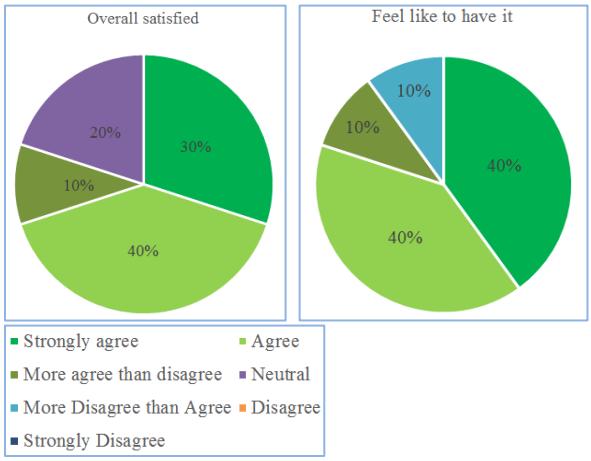


Fig. 22. Satisfaction of VisMarkMap

VI. Discussion

The preliminary observations suggest that VisMarkMap considered as most useful to the major number of users as they were able to do all the desired tasks regarding bookmarking with more organized way in a mind map. The best aspects of this new bookmarking system are mentioned by the users are:

- The visual representation with mind map makes easy the process of remembering the correlations among different bookmarks and causes less mental effort to find out the expected ones. In addition, the five star rating helps them to take quick decision in choosing the most important bookmark.
- The personal and the web notes taking facilities and viewing them along with the zoomed preview give them more flexibility to recall about a bookmark than any other available bookmarking systems.

VII. Conclusion and Future Work

The goal of this work is to find out a new visual bookmarking technique for letting users to organize bookmarks using mind map along with the most expected information, which will help the users to recall easily the previously searched websites' information from the bookmarks and will also minimize the tendency to revisit or research the website using the search engines.

For the future work, the present prototype can be improved further to enhance the scalability and usability. In addition, more experiments with this new system are planned to conduct with more users in the field studies to test with more real time scenarios.

REFERENCES

[1] Zhang,Ping & Galleta,Dennis (2006). Human Computer Interaction and Management Information Systems Foundations.

[2] Alpert,S.R.,& Gruenenberg,K.(2000). Concept mapping with multimedia on the web. Jour-nal of Educational Multimedia and Hypermedia, 9(4), 313-330.

[3] Peter Mika (2007).Social Networks and the Semantic Web, Semantic Web and Beyond.

[4] LjupcoJovanoski, Vladimir Apostolski and DimitirTrajanov (2010). Comparing Social Bookmarking and Tagging Systems: Towards Semantic Sharing Platforms.

[5] Kaasten,Shaun (2000) &Greenberg,Saul. Designing an Integrated Bookmark / History System for Web Browsing. In Proceedings of the Western Computer Graphics Symposium 2000, (Panorama Mountain Village, BC, Canada) , March 26-29, 2000

[6] Cañas, A.J., Leake, D.B., & Wilson, D.C. (1999). Managing, mapping and manipulating con-ceptual knowledge. AAAI Workshop Technical Report WS-99-10: Exploring the synergies of knowledge management & case-based reasoning. Menlo Park, CA: AAAI Press.

[7] John W. Budd (2004). Mind Maps as Classroom Exercises. The Journal of Economic Education , Vol. 35, No. 1 (Winter, 2004), pp. 35-46 Published by: Taylor & Francis, Ltd. Article Stable.

[8] Tanja Keller and Sigmar-Olaf Tergan,Visualizing Knowledge and Information: An Introduction Lecture Notes in Computer Science Volume 3426, 2005, pp 1-23.

[9] Gotz, D. The ScratchPad: sensemaking support for the web. In Proc. Of WWW 2007, 1329-1330.

[10] Levene, M. (2006). An Introduction to Search Engines and Web Navigation. Reading, MA: Addison-Wesley

[11] Tauscher, L. and Greenberg, S., How People Revisit Web Pages: Empirical Findings and Implications for the Design of History Systems. In International Journal of Human-Computer Studies, pages 47(1), 97-138, 1997.

[12] Abrams, D., Baecker, R. and Chignell, M. Information Archiving with Bookmarks: Personal Web Space Construction and Organization. In Proceedings of the ACM/SIGCHI Conference on Human Factors in Computing Systems (CHI'98), pages 18-23, 1998.

[13] Cernea, D., Ebert, A., Kerren, A., & Truderung, I.(2013). WebComets: A Tab-Oriented Approach for Browser History Visualization. GRAPP.

[14] Sweller, J., & Chandler, P. (1994). Why some material is difficult to learn. Cognition and In-struction, 12 (3), 185-233.

[15] Cox,R.(1999).Representation, construction, externalised cognition and individual differences. Learning and Instruction, 9, 343-363.

[16] Larkin,J.H. (1989).Display-based problem solving. In D. Klahr, & K. Kotovsky (Eds.), Complex information processing. The impact of Heribert Simon(pp. 319-342). Hillsdale, NJ: Lawrence Erlbaum Associates.

[17] Larkin,J.H., & Simon, H.A. (1987).Why a diagram is (sometimes) worth 10.000 words. Cognitive Science, 11, 65-100.

[18] Scaife,M., & Rogers, Y. (1996). External cognition: how do graphical representations work? Int. J. Human-Computer Studies, 45, 185-213.

[19] Hix,Deborah and Hartson,H.Rex (1992). Formative Evaluation: Ensuring Usability in User Interfaces. Technical Report TR-92-60, Computer Science, Virginia Polytechnic Institute and State University.

[20] Lund,A.M.(2001).Measuring usability with the USE questionnaire. Usability and User Experience, 8(2), 8

[21] Lewis,J.R.(1995).IBM computer usability satisfaction questionnaires: psychometric evaluation and instructions for use. International Journal of Human-Computer Interaction, 7(1), 57-78.

[22] Brooke,John(1996). SUS-A quick and dirty usability scale. Usability evaluation in industry, 189, 194.

A Sales Forecasting Model in Automotive Industry using Adaptive Neuro-Fuzzy Inference System(Anfis) and Genetic Algorithm(GA)

Amirmahmood Vahabi
Department of IT Management,
Science and Research branch,
Islamic Azad University,
Tehran, Iran

Shahrooz Seyyedi Hosseininia
Department of Industrial Management,
Karaj branch,
Islamic Azad University,
Tehran, Iran

Mahmood Alborzi
Department of IT Management,
Science and Research branch,
Islamic Azad University,
Tehran, Iran

***Abstract*—Nowadays, Sales Forecasting is vital for any business in competitive atmosphere. For an accurate forecasting, correct variables should be considered. In this paper, we address these problems and a technique is proposed which combines two artificial intelligence algorithms in order to forecast future automobile sales in Saipa group which is a leading Automobile manufacturer in Iran. Anfis is used as the base technique which is combined with GA. GA is used in order to tune the Anfis results.**

Furthermore, sales forecasting is succeeded with annual data of years between 1990 and 2016. With this in mind, per capita income, inflation rate, housing, Importation, Currency Rate (USD), loan interest rate and automobile import tariffs are selected as effective variables in the proposed model. Finally, we compare our model with ANN model which is a well-known forecasting model.

Keywords—Sales Forecasting; Adaptive Neuro-fuzzy inference system (Anfis); Genetic Algorithm (GA)

I. INTRODUCTION

Sale is the vital part of any business. Accordingly sales forecasting plays an important role in a business finance planning and is a self-assessment tool for a company. The managers have to keep taking the pulse of their company to know how healthy it is. A sales forecast reports, graphs and analyzes the pulse of the business. It can make the difference between just surviving and being highly successful in business. It is a vital cornerstone of a company's budget. The future direction of the company may rest on the accuracy of sales forecasting.

Nowadays, Automobile industry owns a great place in every company around the globe. As said, sales forecasting helps the company in achieving its goals such as sales revenue increasement, efficiency improvement, customer care, etc. However, still it is one of the hardest fragments of management. [1] [2]

Accurate forecasting allows the firm owners to improve market performance, gain more profit and plan its policies and procedures.

Over the last few decades when dealing with the problems of sales forecasting, traditional time series forecasting methods, such as exponential smoothing, moving average, Box Jenkins ARIMA, and multivariate regressions etc., have been proposed and widely used in practice to account for these patterns, but it always doesn't work when the market fluctuates frequently and at random [3]. Therefore, Research on novel business forecasting techniques have evoked researchers from various disciplines such as computational artificial intelligence.

Automobile market is one of the main industries in Iran which plays a vital role in country's economy. In recent years specially after sanctions which are held in 2011, Iran's automobile sales market was downsized so companies in Iran manufactured much less load than the actual production capacity. Companies' roadmap wasn't ready for this rapid change as forecasting charts wasn't able to predict these circumstances. Consequently company's revenue has dropped. In the other hand, after the joint comprehensive plan of action which was reached agreement in 2015, Iran's market experienced a new era which the market grow rapidly. Accordingly, many foreign manufacturers decided to invest in Iran market. So, in this competitive market, strategy awareness is vital. An accurate Sales forecasting model which will cover annual Sales would be a must have deal for any company which wants to have a share in the newly risen market.

In this study, we proposed a combined methodology for automobile sales forecasting in Iran. The following chapters in this article will be divided into 4 categories. Literature review and also proposed tools will be explained in chapter 2. The proposed model and methodology, also chosen variables will be specified in chapter 3. In chapter 4 results will be shown with charts and graphs. In the end we will conclude our results in chapter 5.

II. LITERATURE REVIEW

Our review of literature specifies that the automobile sales forecasting study contains two types. The first is the introduction of the common forecast models presenting the sales forecast. The second is the set of studies that uses techniques to forecast the sales of automobile industry.

A. Forecasting

Budgeting planning has an important role in any organization. The key element for a useful budgeting process is an accurate sales forecasting. Business forecasting has been continuously been a vital organization capability for both strategic and tactical business planning. [4] With this in mind, improving the quality of forecasts is still an outstanding question which will come to mind. [5] The reason that sales forecasting is particularly important is the effect that it has on many functions of the organization. [6]

Sales forecasting helps managers to make appropriate decisions in uncertain environment. Sales forecasting could be done by linear and non-linear methods. In recent years many researchers tried to develop a model to predict sales. In order to predict sales, researchers tried to use economic indexes to improve model accuracy. In 2002, Kou proposed a model and used ANN so that not only it was able to learn if-then rules, but also it could recognize the fuzzy weights. The main variable they focused on was the impact of advertising on sales [7]. Wang in 2011 proposed a model based on monthly sales in Taiwan. He used indexes such as average earning of employees in industry and services, the oil prices and the superficial measurements of housing starts and building permits, the index of producer's inventory, average monthly overtime in industry and services as the most effective variables on sales forecasting

B. Artificial Intelligence Algorithms

Statistical methods such as ARIMA and linear regression were tools for forecasters in the past. However, in the recent years, with the development of artificial intelligence models, it seems that novel tools such as ANN outperform the results given by mentioned models in the past. [8] [9] [10] [11] [12] [13]

Still, with the nearly accurate results simulated by ANN, results need to be more accurate. Accordingly, with the development of fuzzy systems and the combination with ANN, Anfis was designed which more accurate results where predicted. [14]

Development of heuristic methods, made it simpler to get results in non-exact equations. ANN technique is used in order to train Fuzzy inference systems which concludes in Anfis technique. Moreover, to get more accurate results heuristic models such as Genetic Algorithm could be used in order to tune Anfis results. In the past studies, GA was used in order to tune Anfis results which the output, outperform other compared models. [15] [16]

According to our studies, to the day, combination of Anfis and Ga isn't used in order to forecast future automobile sales in Iran. Therefore, the aim of this paper is the application of Anfis-Ga model in automobile industry sales forecasting.

III. Methodology

A. ANN (Artificial Neural Network)

ANN is a model for data processing which is inspired by biological nervous systems [11]. The connection between elements largely determines the network function in the nature. The key factor in this model is the proliferation of neurons whose synergy results in problem solving. ANN may consists two or more layers. In the complicated models, output of each neuron is the input for other neuron. In general ANN structure is as shown in Fig. 1.

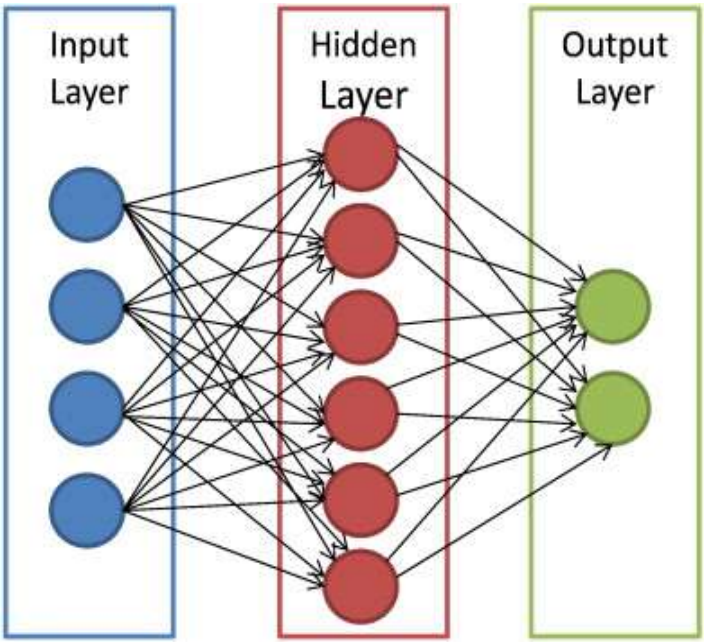


Fig. 1. ANN Structure

General of usage of ANN is for training purposes. A target function is defined and the network adjusts the output by modifying bias and weights of the network. Adjustment may occur by the comparison of output and actual targets. The network learning may continue until the outputs match targets. Normally, many individual input/target pairs are needed for network training. [14]. Usage of neural networks is also for training problems which are not conventional for computers and human knowledge.

B. ANFIS (Adaptive Neuro Fuzzy Inference System)

Anfis technique was first suggested in 1993 by Jang. [17]. Anfis is a hybrid model based in Takagi-Sugeno fuzzy inference system. One of the obstacles of the Fis technique is the extraction of fuzzy rules which when the inputs and outputs or the membership functions are occurred with multiplicity, rules extraction may face more challenging and time consuming, sometimes impossible even for expert knowledge. ANN learning technique is used in order to extract fuzzy rules for the Fuzzy inference system from the past data. The proposed technique is known as Anfis.

For simpler explanation, a fuzzy inference system with 2 inputs(x and y) and one output (z) is considered. For a Sugeno model to organized a fuzzy if-then rule is needed which is expressed as

$$If\ x\ is\ A_1 and\ y\ is\ B_1\ then\ f_1 = p_1 x + q_1 x + r_1 \quad (1)$$

Where p, r, and q are linear output parameters. The architecture of an Anfis consisted of 2 inputs and one output using 5 layers which is shown in Fig 2.

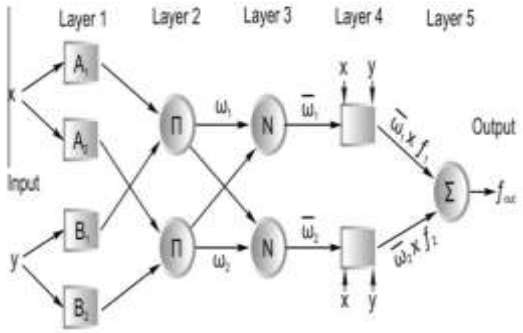


Fig. 2. Anfis structure of two inputs and two rules

Layer-1:

$$O_{1,i} = \mu_{Ai}(x), for\ i = 1,2,3\ O_{1,i} = \mu_{Bi-3}(y), for\ i = 4,5,6 \quad (2)$$

Where x and y are node i inputs, whereas A_i and B_i are inputs' linguistic labels. To simplify, $O_{1,i}$ is membership function for A_i and B_i. $\mu_{Ai}(x)$ and $\mu_{Bi}(y)$ are bell-shaped with maximum number of 1 and minimum equal to 0, as follows:

$$\mu_{Ai}(x), \mu_{Bi-3}(y) = \exp\left(\left(-\frac{(x_i - c_i)}{(a_i)}\right)^2\right) \quad (3)$$

Where a_i and c_i is the parameter set.

Layer-2: Each node in this layer multiplies incoming inputs and send the result as output. The result represents the firing strength of the node.

$$O_{2,i} = w_i = \mu_{Ai}(x).\mu_{Bi-3}(y), \quad i = 1,2,3,\dots,9 \quad (4)$$

Layer-3: In this node the ratio of the ith rules firing strength is the sum of all rule's firing strengths is calculated.

$$O_{3,i} = \bar{w}_i = \frac{w_i}{w_1 + w_2 + \cdots + w_9}, i = 1,2,3,\dots,9 \quad (5)$$

Layer-4: Each node I is a square node with a node function in this layer.

$$O_{4,i} = \bar{w}_i.f_i = w_i.(p_i x + q_i x + r_i), \quad i = 1,2,3,\dots,9 \quad (6)$$

Layer-5: In this layer, the single node is a circle node labeled $\sum$ which computes the total output as the summation of all incoming signals:

$$O_{5,j} = overall\ output = \sum_i \bar{w}_i f_i = \frac{\sum_i \bar{w}_i f_i}{\sum_i w_i} \quad (7)$$

C. GA (Genetic Algorithm)

The Genetic Algorithms is categorized as a heuristic method, which is inspired by natural genetics. Mechanism is using random selections in order to improve the final result. [19] Random population is created and the use of crossover and mutation on the parent chromosomes which produces offspring that have both parents' advantages is occurred in this model. GA operation is as follows:

a) Random individuals are created which fills the initial population. All parameters could be set due to problem.

b) A fitness function must be set in order to rank the created chromosomes. Each individual is evaluated and prioritized based on the fitness measure.

c) If the termination criteria is reached, the best chromosome will be returned as the solution

d) If not, based on problem parameters, selected number of individuals from the initial population would be chosen and the genetic operations (crossover, mutation) will be applied. Newly created population would be evaluated by fitness function and all individuals would be combined and prioritized again based on fitness measures.

e) If still the termination criteria is not met, actions from step 2 will repeat until the termination criterion is satisfied. Each iteration is called generation[21-22]

In the proposed model, GA is used in order to tune the Anfis output. The elements of the Fis which is obtained from Anfis output would be extracted and would be set as chromosome model. Random individuals are created based on the model and also the number range of the genes is set by the problem parameters.

IV. Research Design and Experiments

This paper presents an automobile sales forecasting model based on Anfis and GA. The presented model is tested by Saipa group which is a leading automobile manufacturing company in Iran. The objective of the system is to forecast future annual sales.

The methodology for the steps of training process of the system is explained on Fig. 3.

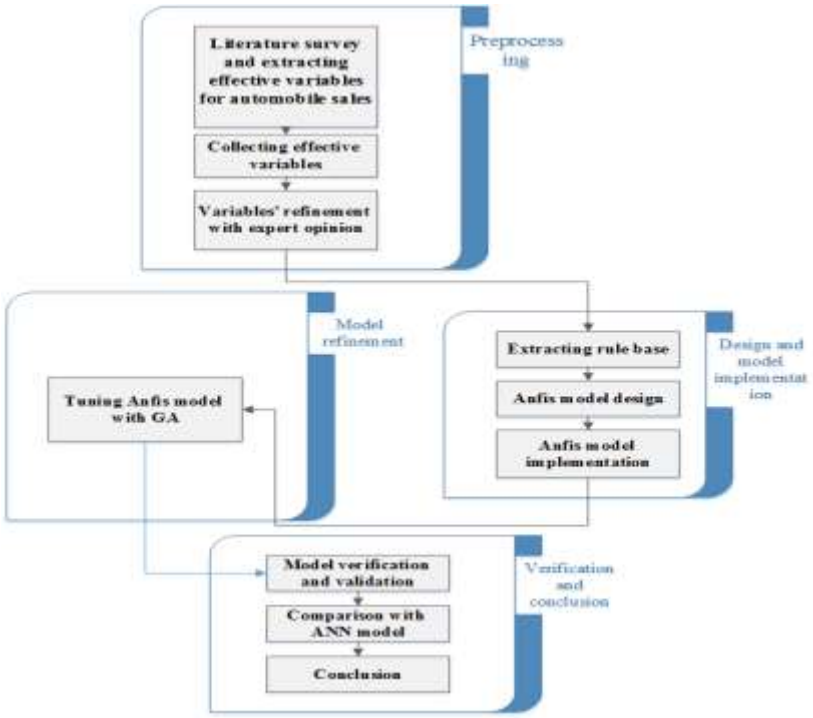


Fig. 3. Model steps

Proposed model is consisted of 4 major steps which are: 1.Preprocessing 2.Model design and implementation 3.model refinement 4. Verification and conclusion

A. Preprocessing

From the previous studies the variables which are more effective on automobile sales are extracted. Variables such as population size, unemployment rate, exchange rates against US dollar, production index, real customs-cleared exports, the sales of manufacturing, the sales index of wholesale, the total power consumption, Consumer Price Index, Unemployment Rate, Gas Prices, Housing Starts, Gross Domestic Product, Inflations rate, Base lending rate, stock index, etc. where used in previous studies which are selected and are forwarded to next step. [18] [19] [20] [21]

Additionally, Variables per capita income, inflation rate, housing starts value, Importations value, Currency Rate (USD), loan interest rate and automobile import tariff are chosen as input for the system by experts' opinion based on checking correlations of each variable and market experience.

In this research annual data of each variable is collected from 1990 until 2005. A yearly sale of automobiles is also collected and all the data is integrated in one table and is ready for training. Sample of prepared data can be seen in Table I.

TABLE I. SAMPLE PREPARED DATA

year	currency rate against USD	inflation rate	per capita Income	loan interest rate	import tariff	importations value	housing starts value	Total Sales
1990	5388.839	17.40%	2228114	10	0.35	22013392857143	4766964285714	10894
1991	5788.115	9.00%	2795684	18	0.35	37053278688525	7992213114754	20878
1992	4814.237	20.70%	3183551	18	0.35	55372881355932	9982372881356	30859
1993	4081.744	24.40%	3094298	18	0.35	53324250681199	10391825613079	18536
1994	4013.333	22.90%	3869217	21	1	48657777777778	9274000000000	13895
1995	4326.765	35.20%	3739266	21	1	30568144499179	9141215106732	15642

B. Model design and implementation

System is trained using Anfis technique. Input dataset is divided in two groups. 70% is chosen for training and remaining 30% is picked as test data which is used for validation. Fitness functions RMSE and R^2 are used for model comparison and results evaluation.

In order to train the Anfis with most valuable data, shuffling technique used. All data is shuffled in rows before selecting train data and test data. With this method train data is chosen randomly from different years. [21]

Fuzzy C-means method is used for initial Fis generation. Default parameters are used for Anfis procedure in which initial step size is set as 0.01. Decrease rate and increase rate for Anfis parameters are set as 0.9 and 1.1.

C. Model refinement

In this process, GA is used as a refinement for Anfis output results. The output Fis which is casted by Anfis training, is extracted to its core elements and labeled as P which gives us a 1 * N matrix. The extracted matrix's number of columns is modelled and a raw matrix with N columns is initialized which is labeled as w. Our purpose is to generate an optimum w matrix with the operations of GA which after multiplying this matrix to our base Fis elements which is P, would give us a new matrix with N columns and 1 row named as P'. P' could be transferred to a new Fis which can be evaluated by inputting data. The optimum P' which is generated after termination of GA is the refined Fis which is the output of the system.

$$P' = P \,.\, w \tag{8}$$

GA parameters for tuning Anfis are set as below:

Population size: 50

Max iteration: 200

Crossover percentage: 80%

Mutation percentage: 40%

Selection Pressure: 8

D. Verification and conclusion

As stated on previous sections, 30% of data is selected as test data and is as test data isn't involved in any part of the training process, it is used as model validation. RMSE and R^2 are designated as indicators for validation.

The Root Mean Square Error (**RMSE**) (also called the root mean square deviation, RMSD) is a commonly used measure of the difference amongst values projected by a model and the values actually observed from the modelled environment.

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(X_{obsi} - X_{model,i})^2}{n}} \tag{9}$$

Where X_{obs} is observed values and X_{model} is model evaluated values at time/place *i*.

In statistics, the coefficient of determination, denoted R^2 or r^2 and pronounced "R squared", is a number that specifies the variance in the independent variable from the dependent variable which is predictable. [23]

Based on the proportion of total variation of outcomes explained by the model, R^2 provides a measure of how well observed outcomes are simulated by the model, [24]

$$R^2 \equiv 1 - \frac{\sum_i (y_i - f_i)^2}{\sum_i (y_i - \bar{y})^2} \tag{10}$$

In the next step, result of the model is compared with results derived from ANN model.

V. Empirical Results

After implementing above-mentioned data with our model, annual forecasting evaluations for automobiles based on previous yearly sales and other factors during 1990-2016 are summarized in below.

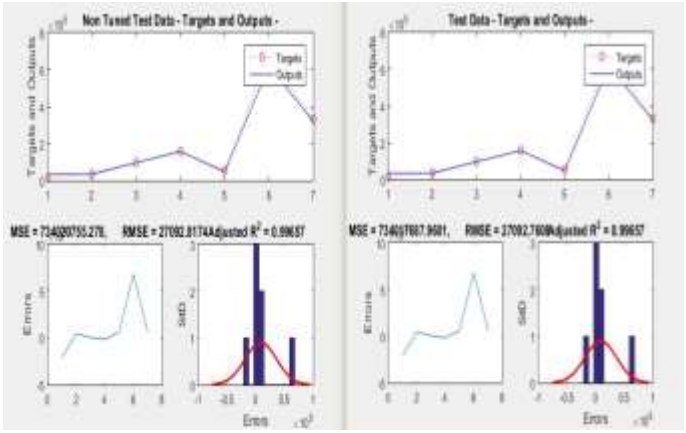


Fig. 4. Anfis-GA and Anfis model results comparison

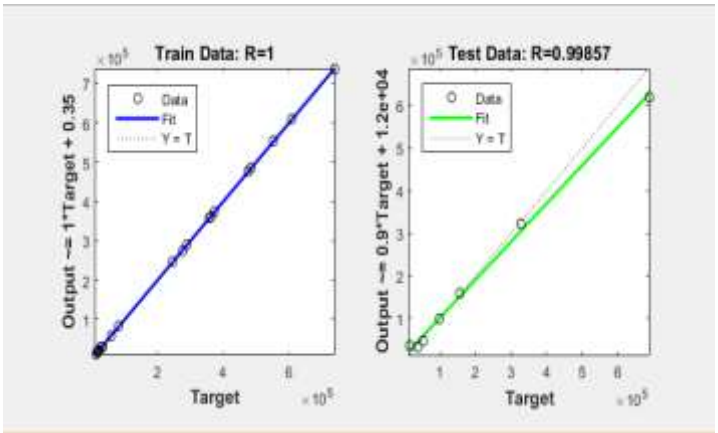


Fig. 5. Anfis results regression plot

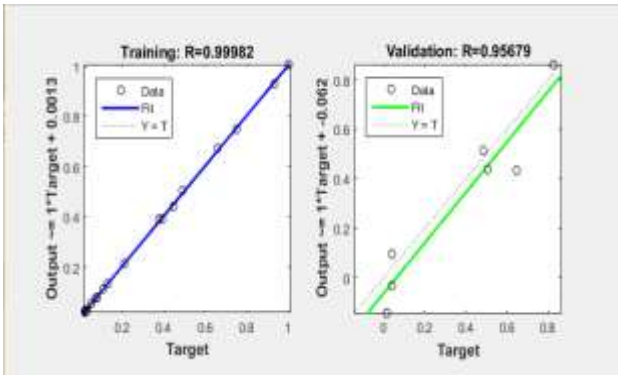


Fig. 6. ANN results regression plot

TABLE II. ANNUAL SALES DATA COMPARISON

Year	Actual	Anfis	Anfis-GA	ANN
1990	10894	10894	10894	12776.3
1991	20878	20878	20878	20657.1
1992	30859	30858.4	30858.5	30837.6
1993	18536	18536.6	18536.6	17146.9
1994	13895	34936.7	34936.8	13506.6
1995	15642	15642	15642	14721.4
1996	18750	18750	18750	17774.1
1997	28758	28759.7	28759.7	29158
1998	38494	33779.1	33779.1	9567.51
1999	53210	46991.6	46991.6	17879.9
2000	57856	57854.3	57854.4	57367.9
2001	81980	81980.5	81980.2	81496.6
2002	97557	97093.3	97093	96303.4
2003	155880	157490	157491	204307
2004	245431	245430	245431	241285
2005	289234	289235	289234	282584
2006	372387	372393	372393	364094
2007	474855	474849	474849	468972
2008	484945	484945	484945	488537
2009	553502	553504	553504	374864
2010	608914	608914	608914	553111
2011	687028	619231	619231	570481
2012	736614	736611	736612	714066
2013	361486	361484	361484	373172
2014	275672	275674	275674	284691
2015	357441	357441	357441	358789
2016	328170	322231	322232	327672

TABLE III. MODEL RESULTS COMPARISON

Model	RMSE	R^2
Anfis	27092.81	99.66%
Anfis-GA	27092.76	99.66%
ANN	59436.75	95%

Results obtained from Table III indicates the point that Anfis-GA model has reduced RMSE and succeeded in model tuning, although results improvement is not tangible and R^2 outputs in test data hasn't changed.

VI. CONCLUSION

This paper proposes a methodology for forecasting automobiles sales data in a manufacturing company in Iran using combination of Anfis and GA. The output system is capable of running for future forecasts. The proposed methodology has been evaluated in an automobile manufacturing company using real consumption data.

Anfis technique is used for training system as GA is used in order to tune Anfis results. Output results have been compared with ANN model and it is seen that Anfis results has outperformed ANN. With the comparison of Anfis model with Anfis-GA tuned outputs, results show that GA operates effective tuning in training procedure. Evaluated results in test data indicated the point that GA tuning may improve results. Although, improvement isn't sensed in R^2 results.

Moreover, an important advantage of this methodology is that it can be updated in intervals with new data.

To conclude, in this paper our Anfis-Ga methodology is used for forecasting annual sales data in an automobile manufacturer in Iran. Results show that although using GA for tuning results outperform other results' evaluated by other models such as ANN and Anfis.

For future studies, other heuristic and meta-heuristic methods could be used for tuning Anfis results. Also Heuristic methods such as GA can be used as selection method for input variables before training with Anfis.

REFERENCES

[1] P. Chang, C. Liu and C. Fan, "Data clustering and fuzzy neural network for sales forecasting: A case study in printed circuit board industry," Knowledge-Based Systems, vol. 22, no. 5, pp. 344-355, 2009.

[2] A.-E. S. A. and F. Mannering, "Forecasting automobile demand for economics in transition, a dynamic simultaneous system approach," Transportation Planning and Technology, vol. 25, pp. 311-331, 2002.

[3] M. Lawrence and M. O'Connor, "Sales Forecasting Updates: How Good Are They in Practice?," International Journal of Forecasting, vol. 16, no. 3, pp. 369-382, 2000.

[4] R. Filders and R. Hastings, "The Organization and Improvement of Market Forecasting," Journal of Operation Research Society, vol. 45, no. 1, pp. 1-16, 1994.

[5] C. W. J. Granger and M. O'Connor, "Sales Forecasting Updates: How Good Are They in Practice?," International Journal of Forecasting, vol. 16, no. 3, pp. 369-382, 2000.

[6] J. T. Mentzer and C. C. Bienstock, Sales Forecasting Management: Understanding the Techniques, Systems and Management of the Sales Forecasting Process, Thousand Oaks, CA: SAGE Publications, Inc, 1998.

[7] R. Kou, P. Wu and C. Wang, "An intelligent sales forecasting system through integration of artificial neural networks and fuzzy neural networks with fuzzy weight elimination," Neural Networks, vol. 15, pp. 909-925, 2002.

[8] T. Kimoto, K. Asakawa, M. Yoda and M. Takeoka, "Stock market prediction system with modular neural networks," in Proceedings of the international joint conference on neural networks, San Diego, 1990.

[9] T. W. S. Chow and C. T. Leung, "Nonlinear autoregressive integrated neural network model for short-term load forecasting," IEE Proceeding Online, vol. 19960600, pp. 500-506, 1996.

[10] R. Law and N. Au, "A neural network model to forecast Japanese demand for travel to Hong Kong," Tourism Management, vol. 20, pp. 89-97, 1999.

[11] J. T. Luxh, J. O. Riis and B. Stensballe, "A hybrid econometric-neural network modeling approach for sales forecasting," The International Journal of Production Economics, vol. 43, pp. 175-192, 1996.

[12] A. S. Tawfiq and E. A. Ibrahim, "Artificial neural networks as applied to long-term demand forecasting," Artificial Intelligence in Engineering, vol. 13, pp. 189-197, 1999.

[13] F. M. Thiesing, U. Middelberg and O. Vornberger, "A neural network approach for predicting the sale of articles in supermarkets.," Third European Congress on Intelligent Techniques and Soft Computing, pp. 28-31, 1995.

[14] A. Dwivedi, M. Niranjan and K. Sahu, "A Business Intelligence Technique for Forecasting the Automobile sales using Adaptive Intelligent Systems," International Journal of Computer Applications, vol. 74, no. 9, pp. 7-13, 2013.

[15] L.-Y. Wei, "A GA-weighted ANFIS model based on multiple stock market volatility causality for TAIEX forecasting," Applied Soft Computing, vol. 13, pp. 911-920, 2013.

[16] K. Kampouropoulos, F. Andrade, J. Cardenas and J. Romerar, "A Methodology for Enery Prediction and Optimization of a System based on the Energy Hub Concept using Particle Swarms," The Annual Seminar in Automation, Industrial Electronics and Instrumentation, 2012.

[17] J. Jang, "ANFIS: adaptive-network-based fuzzy inference system," IEEE Transactions on Systems, Man, and Cybernetics, vol. 23, no. 3, pp. 665-685, 1993.

[18] A. Sa-ngasoongsong, S. T. Bukkapatnam, J. Kim, P. S.lyer and R. Suresh, "Multi-step sales forecasting in automotive industry based on structural relationship identification," International Journal of Production Economics, vol. 140, no. 2, pp. 875-887, 2015.

[19] F.-K. Wang, K.-K. Chang and C.-W. Tzeng, "Using adaptive network-based fuzzy inference system to forecast automobile sales," Expert Systems with Applications, vol. 38, pp. 10587-10593, 2011.

[20] F. Muhammad, M. Y. M. Hussin and A. A. Razak, "Automobile Sales and Macroeconomic Variables: A Pooled Mean Group Analysis for Asean Countries," IOSR Journal of Business and Management, pp. 15-21, 2012.

[21] F.-C. Yuan, "Parameters Optimization Using Genetic Algorithms in Support Vector Regression for Sales Volume Forecasting," Applied Mathematics, vol. 3, pp. 1480-1486, 2012.

[22] M. Jalali-Heravi and A. Kyani, "Comparison of Shuffling-Adaptive Neuro Fuzzy Inference System (Shuffling-ANFIS) with Conventional ANFIS as Feature Selection Methods for Nonlinear Systems," vol. 26, no. 10, p. 1046 – 1059, 2007.

[23] Stat Trek, "Stat Trek website," 2016. [Online]. Available: http://stattrek.com/.

[24] S. A. Glantz and B. K. Slinker, Primer of Applied Regression and Analysis of Variance., McGraw-Hill. ISBN 0-07-023407-8, 1990.

Japanese Dairy Cattle Productivity Analysis using Bayesian Network Model (BNM)

Iqbal Ahmed
Graduate School of Science and Engineering
Saga University, Saga, Japan

Osamu Fukuda
Graduate School of Science and Engineering
Saga University, Saga, Japan

Hiroshi Okumura
Graduate School of Science and Engineering
Saga University, Japan

Kenji Endo
Morinaga Dairy Service Co. Ltd.
1-159 Toyoharaotsu, Nasugun Nasumachi,
Tochigi 329-3224, Japan

Kohei Arai
Graduate School of Science and Engineering
Saga University, Japan

Kenichi Yamashita
Advanced Manufacturing Research Institute, The National Institute of Advanced Industrial Science and Technology(AIST), Tosu 841-0052, Japan

***Abstract*—Japanese Dairy Cattle Productivity Analysis is carried out based on Bayesian Network Model (BNM). Through the experiment with 280 Japanese anestrus Holstein dairy cow, it is found that the estimation for finding out the presence of estrous cycle using BNM represents almost 55% accuracy while considering all samples. On the contrary, almost 73% accurate estimation could be achieved while using suspended likelihood in sample datasets. Moreover, while the proposed BNM model has more confidence than the estimation accuracy lies in between 93 to 100%. In addition, this research also reveals the optimum factors to find out the presence of estrous cycle among the 270 individual dairy cows. The objective estimation methods using BNM definitely lead a unique idea to overcome the error of subjective estimation of having estrous cycle among these Japanese dairy cattle.**

Keywords*—Bayesian Network Model; BCS; Postpartum Interval; Parity Number; Estrous Cycle; Cattle Productivity

I. Introduction

Dairy cattle productivity largely depends on pure and more accurate understanding of the presence of estrous cycle. The subjective methods of finding estrous cycle in cows, such as ultrasound image analysis by an experienced inspector could jeopardize the farm productivity. The Bayesian Network Model (BNM) with the inclusion of Body Condition Parameter (BCS), Postpartum Interval (PPI) and Parity number could be used to overcome the error in subjective estimation of estrous cycle presence in cows. This research reveals that the approach of using BNM with other parameters, exhibit more objectively accurate estimation of the presence of estrous cycle in the industry and thereby, helping to the farm management to design proper estrous synchronization protocol. The Body Condition Score: BCS, Postpartum Interval: PPI and Parity Number have taken into consideration for designing the proposed BNM.

The paper describes research background first followed by introducing of BNM. Then experimental procedure and results is described followed by conclusion with some discussions.

II. Research Background

A sound understanding of the presence or absence of estrous cycle allows cattle producers to troubleshoot reproductive problems in their farms. These understanding are also important when using estrous synchronization and other reproductive technologies in dairy industry. The estrous cycle of cattle is the period from one estrus (heat, phase of sexual receptivity) to the next estrus. For the cow and heifer, this period averages 21 days, with a typical range of 18 to 24 days in length [1, 2, 3, 4]. The reproductive function of a cow or heifer is characterized by whether she displays normal estrous cycles or not. Many factors have potential influences on the presence or absence of estrous cycle of heifers and researchers already focused on it elaborately [5, 6, 7, 8, 9, 10, 11, 12]. Body condition Score (BCS), Days after childbirth and or Postpartum Interval(PPI), parity number, ovarian characteristics, uterine blood flow, progesterone level(P4), climate and nutritional factors are mostly discovered influential factors in this arena [13, 14, 15, 16]. Moreover, the various species of heifers and different country's environmental condition could play vital roles in this case. However, discovering the presence of estrous cycle would definitely an important indication for reproductive management in the cow herd. This would help the herd management to synchronize the estrous and thus have more chance to make cow pregnant with the help of artificial insemination or even in a natural way [17, 18]. It is evident from [17] that, estrous synchronization protocol assists to get higher pregnancy rate in many countries (Fig. 1). However, finding out the presence of estrous cycle and pregnancy investigation in cattle usually done by ultrasound image analysis by an experienced inspector and some regular data analysis tools [18, 19, 20, 21, 22]. The usage of ultrasound image analysis for cattle is still very much subjective and expensive to some extent. Moreover, much experiences and skills are required for interpreting ultrasound image analysis to identify estrous cycle properly.

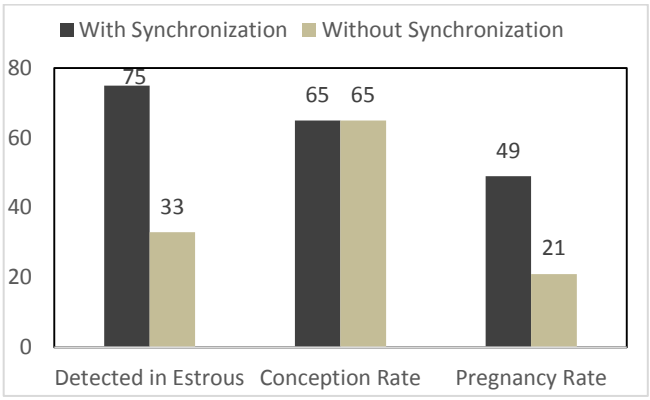


Fig. 1. Effect of Estrous Synchronization Protocol, adopted from [17]

Among the many factors of cattle productivity influence, the most influential one is the BCS, which is defined as "an effective management tool to estimate the energy reserves of a cow" [13, 23, 24] and most widely used for herd management. There are many identified systems for measuring BCS, which varies according to different countries [13, 23]. Using BCS to evaluate cattle does not require any special equipment and can be conducted anytime during the year. Poor body condition is associated with reduced income per cow, increased postpartum interval, increased dystocia, and lower weaning weight. The most common and widely used (USA and Japan) BCS scale ranges from 1 to 5 with 0.25 increments [24]. Though BCS measured subjectively and its reliability is questioned, it is also evident that BCS has relationship with many other factors of bovine, such as postpartum interval, parity, and etc. [5, 11, 15, 24]. However, this investigation focused on three influential factors (BCS, postpartum interval, and parity) for understanding the presence and absence of estrous cycle using a new unique Bayesian Network Model (BNM). In total, 280 different Japanese Holstein cows observing with their BCS (2.0 to 3.25), postpartum interval and parity numbers to discover the ideal timing for artificial insemination to make them pregnant. It is also important to mention that, all these 280 samples found anestrus in their farm. The aim of this study is to find out the optimum factors to have an estrous cycle of bovine using Bayesian network model analysis in Japanese dairy industries. It is clear from National Livestock Breeding Center (NLBC, Japan) that, the overall conception rate of live beef and dairy cattle is decreasing in last 20 years in Japan (Fig. 2) [25]. Moreover, the findings of Bayesian network analysis could use for designing estrous synchronization protocol to improve cattle productivity and herd management. Moreover, using BNM analysis would assist the farm management to find out the presence of estrous cycle more objectively and in an accurate way.

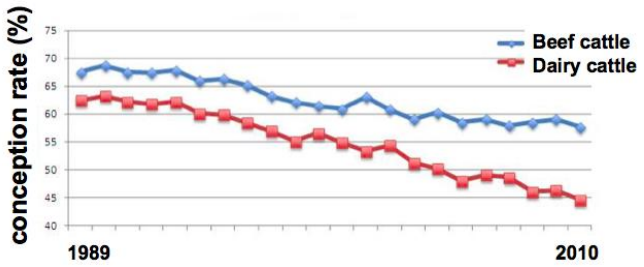


Fig. 2. The decreasing conception rate in Japan, partial adoption from NLBC, Japan

Using BNM for analyzing the presence or absence of estrous cycle is the most unique work in this arena. In addition, the Bayesian results deliver higher accuracy to find out the estrous cycle in relationships with BCS, Postpartum interval and number of parity. The rest of the paper is organized as follows, next the Bayesian network model section describes the proposed designed Bayesian network model for identifying estrous cycle using the data sets. Experiment section will introduce the overall experimental steps, data collection methods, and conditions in detail. The analytical results and its interpretation will present in Results and Discussion sub section. Finally, the paper concludes with future plans of this research and few challenges.

III. Identifying Estrous Cycle using Bayesian Network Model (BNM)

Fig.3 depicts the Bayesian network used in this investigation. Bayesian network is represented using the directed graph. The parent node indicates the cause, and the child node indicates the result. The proposed method uses three kind of information. Each node indicates the valuable such as BCS, (PPI)/Days after childbirth, Parity number, and Estrous cycle. The details are explained in the following fig. 3.

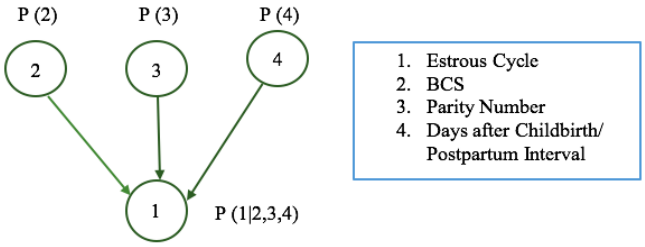


Fig. 3. The proposed Bayesian Network tree structure for identifying estrous cycle in Cattle

According to Bayes rules, posterior probability of having estrous cycle or not is designed with each individual parameter's prior and conditional probability. Here, The Bayes equation is used for each parameter as follows,

$$P(estrous\ cycle|BCS) = \frac{P(estrous\ cycle)P(BCS|estrous\ cycle)}{P(BCS)} \quad (1)$$

where, $P(estrous\ cycle)$ = Prior probability of having estrus or not; $P(BCS|estrous\ cyle)$ = conditional probability of having estrus or not based on BCS information; $P(BCS)$ = probability of BCS evidence to find the posterior probability of having estrous cycle. Finally, the total posterior probability of the presence or absence of estrous cycle was determined under three evidences (BCS, PPI, Parity) by following (1).

A. Body Condition Scoring (BCS)

The research reveals to include BCS while considering the estrous cycle identification. The BCS is the most significant influential factors in bovine productivity. An organized process for determining BCS was created at the University of Pennsylvania to help achieve consistency and repeatability in BCS. This system finds its accuracy toward the mid-range scores (2.50 to 4.00), which includes most cattle in this investigation. This mid-range is the most critical for making farm management decisions and most influential for the farm nutritionist. The BCS outside this range indicate significant problems and varies significantly with respect to each individual inspector/observer. This research considering $BCS_{4.0}$ methods (quarter-point increase) in 280 individual cattle of Morinaga Dairy Service Co. Ltd. (MDS), Japan and the following table describes the meaning of BCS scale. The $BCS_{4.0}$ method (0.25 increase) have good repeatability across and within observers including simplified body scoring as well as have higher value as a diagnostic test [24]. The BCS process represents the observer's view into the certain anatomical sites for each cow's pelvic, loin areas, pin and hook bones, and etc. Next Table I briefly elaborates the observing BCS of 280 individual cows from the dairy farm of Iwate Prefecture, Japan under MDS cooperation.

TABLE I. BCS AND IT'S GENERAL MEANING FOR 280 SAMPLE COW

BCS	Meaning (in general)
2.25	No fat pads on pin and hook bones- angular shape
2.5	Palpable fat pads on pin and hook bones- angular shape
2.75	Pin bones- round shape and hook bones- angular shape with less fat pads
3.0	Fat pads on pin and hook bones- round shape
3.25	Visible fat pads on pin and hook bones- round shape

B. Postpartum Interval(PPI)/Days after Childbirth

Each cow goes through a period of temporary infertility known as postpartum anestrus. Usually, cattle do not have estrous cycle during this period. The common term associated with this is postpartum interval (PPI), which is the duration from calving to the subsequent conception again. Several factors affect the postpartum interval of cows such as BCS, age and genetics. This research also includes PPI or days after calving parameter to evaluate the finding of estrous cycle in cows. It is also evident that BCS is affecting the PPI in beef and dairy cattle [5, 14, 15, 16, 26]. Therefore, this research introduces PPI as an individual parameter in the proposed BNM, which assists to find more accurate estimation of estrous cycle with sample data. The following table II describes and categories the PPI of 280 individual cows. In total, the proposed BNM consider 9 groups with one month (30 days) interval.

C. Parity Number/Number of calves

Parity is another important parameter in the proposed BNM. The total number of calves plays vital role in this investigation as it affects the probability of getting pregnant in the next subsequent time. The parity number is calculated without considering its first birth. In this investigation, among the 280 individual Holstein cattle, the highest number of parity for each cow is 9 and the lowest is 1. Usually, the cow with higher parity might have less chance to resume their estrous cycle whereas the cow with parity range of 1 to 4 might have higher possibility to continue their estrous cycle on time.

TABLE II. DAYS AFTER CHILDBITH/PPI & IT'S GROUP FOR BNM

Days after Calving/PPI	Grouping
31-60 days	1
61-90 days	2
91-120 days	3
121-150 days	4
151-180 days	5
181-210 days	6
211-240 days	7
241-270 days	8
>271 days	9

IV. EXPERIMENTAL SETUPS

To evaluate the finding of estrous cycle while considering BCS, PPI and Parity number, this research proposed a unique Bayesian Network Model (Fig.3). The overall experimental steps illustrate in the next Fig.4. Each individual cattle are identified with a unique number in the farm and then observed by an experienced inspector. The required BCS, PPI, Parity number and estrous cycle related data were collected. These sample data were then processed according to the proposed BNM to learn the system. The BCS measurement were carried out with well know $BCS_{4.0}$ system and the PPI interval were categorized into 9 groups (Table I and II). Then, all these four individual parameters for each 280 individual data were feed into the proposed Bayesian model. The learning and validation of the model is described briefly in the Results and Discussion sub section.

A. Conditions

The followings are some of the important condition to mention during this research. All these 280 individual sample data were collected from the dairy farm of Iwate Prefecture with the cooperation of Morinaga Dairy Service (MDS) Co. Ltd., Japan. The BCS were observed in accordance with the UV method of Ferguson [24] by an experienced animal scientist of MDS. The PPI, Parity and other related information is collected from MDS. All of these 280 individual cattle were Japanese Holstein breed, which were found anestrus due to some health hazards in its own farm. The overall investigation for all these problematic dairy cow is under observation of MDS. The Bayonet 6 software, developed by AIST, Japan was used to design the proposed Bayesian network model. The final analytical results represented by using JMP data analytical tool, developed by SAS Institute Inc.

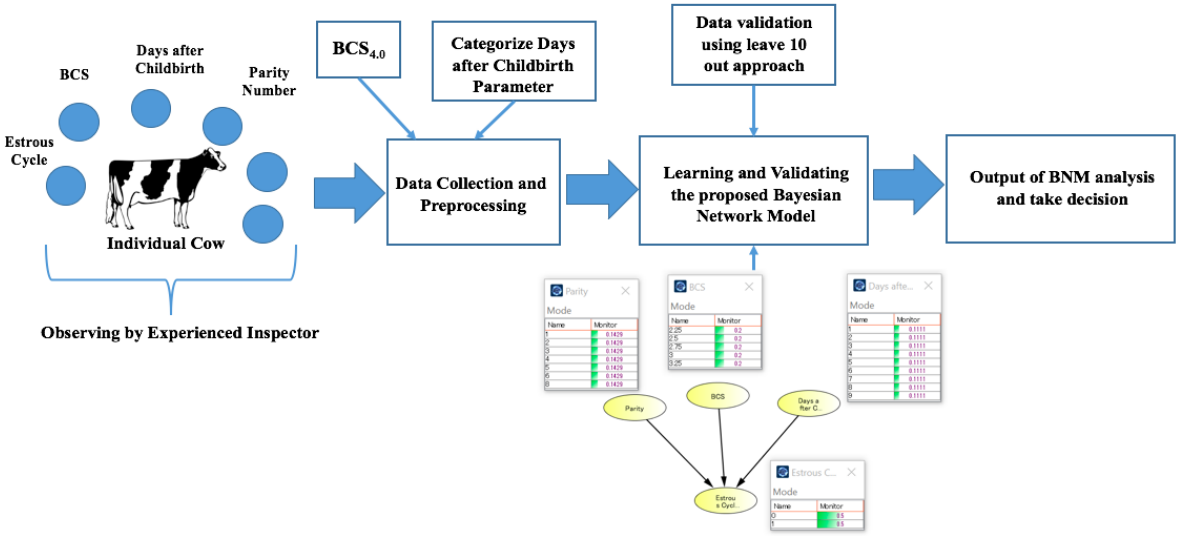


Fig. 4. The overall investigation methods using BNM

B. Results and Discussions

Bayesian network model assisted to visualize the changes of posterior probability as the evidence increases and thus assists to improve the accurate findings of estrous cycles with other methods. The approach of using BNM with the inclusion of BCS, PPI and parity parameters overlooked all previous estimation error of finding estrous cycle. The learning of Bayesian network includes 270 individual data and the rest of 10 data used for test purpose. The 10 data leave out approach improves the accuracy of the model and therefore 27 individual sets of learning and testing data sets used to validate the proposed model. Moreover, when the proposed model acquired higher confidence (less entropy value), the estimation accuracy for the presence of estrous cycle lies in between 93 to 100%.

There were 270 samples used to learn the model and the learning is based on Greedy search algorithm. As a measurement criterion for the appropriateness of a graph structure, information criteria AIC is used. The Bayesian tree was formed by using the estimation of BCS, PPI and Parity number of each cow. Finally, the posterior probability was calculated under these three evidence by following (1). The overall optimum factors for the presence of estrous cycle according to BCS are illustrated in next table III. The fact of the table is the general output of the proposed BNM with learning datasets.

It is evident from many researches that, BCS plays most vital role for affecting all other individual parameters. Therefore, the table III focused only on BCS mostly. The posterior probability for the presence of estrous cycle according to BCS, PPI (1 to 4 groups) and Parity (1 to 4) is measured with 270 sample data. When the BCS is 2.75, the PPI is 31-60 days (group 1) and the parity is 1, the probability of having estrous cycle is 80%. At the same time, when the BCS is 2.5, PPI is 91-120 days (group 2) and Parity is 3, the probability of the presence of estrous cycle is same (80%). The benefit of using Bayesian model is to find out easily these kind of many more relationships in the productivity management. It is now also clear from the table that, BCS is actually not only the significant factor affecting the presence or absence of timely estrous cycle in cattle. Next, the analysis of BNM with 27 individual test datasets is shown in fig. 5.

The overall highest estimation accuracy for finding the presence of estrous cycle based on proposed BNM is 93% and the average accuracy for all data set is almost 55% and lowest average accuracy is 50%, while the log-likelihood is more than 0.7. Using the suspended rule on average likelihood, this research discovered the average estimation accuracy of finding estrous cycle, which represents in next fig. 6.

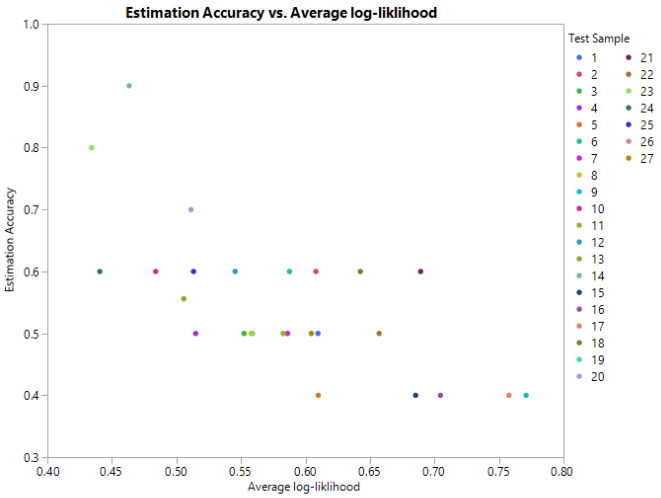


Fig. 5. Estimation accuracy of individual data sets

It is clear from fig. 6 that, when the likelihood is <0.5, the highest average estimation accuracy for the presence of estrous cycle is almost 73% (total 40 cattle) and lowest is 50% when the likelihood is considering <0.7. On the contrary, most of the dairy cows lies within the likelihood of <0.6 (130 + 40 =170 cattle) and the average estimation accuracy for considering all cattle is 55%. However, the research

discovered that, the sample data might not be enough to satisfy the proposed Bayesian model. Therefore, the entropy of proposed model's outputs was calculated to achieve reliable discrimination and use it for discrimination-suspension rule [27]. Entropy indicates or interprets as the risk of incorrect discrimination and if entropy exceeds some predefined discrimination threshold, then the discrimination could be suspended. The following equation used to calculate the entropy between two states of estrous cycle in the designed model.

$$entropy = -\sum_{i=1}^{2} P_i \, log P_i \quad (2)$$

where, $P_i =$ results of posterior probability for the presence (1) or absence (0) of estrous cycle. The higher entropy means the designed network model is ambiguous and less entropy derive the more confident model.

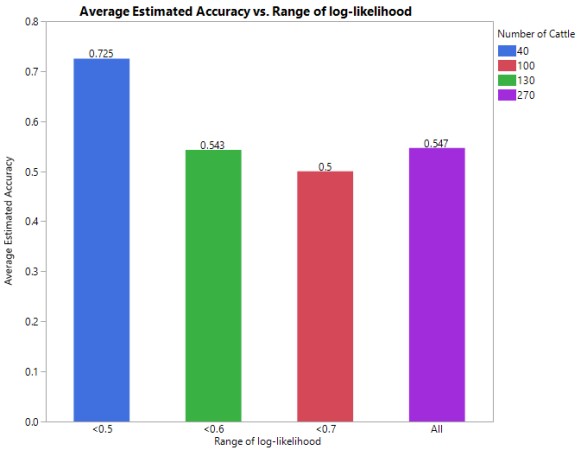


Fig. 6. Average accuracy vs suspended log likelihood based on total number of cattle

TABLE III. THE PROBABILITY OF HAVING ESTROUS CYCLE IN 280 COWS ACCORDING TO BCS, PPI & PARITY

BCS	PPI (according to group)	Parity (parity 1 to 4)	Presence of Estrous cycle (%)
2.25	1 (31-60 days)	1	13
	2 (61-90 days)	2	20
	3 (91-120 days)	3	60
	4 (121-150 days)	4	50
2.5	1 (31-60 days)	1	44
	2 (61-90 days)	2	65
	3 (91-120 days)	3	80
	4 (121-150 days)	4	75
2.75	1 (31-60 days)	1	80
	2 (61-90 days)	2	71
	3 (91-120 days)	3	71
	4 (121-150 days)	4	50
3.0	1 (31-60 days)	1	50
	2 (61-90 days)	2	75
	3 (91-120 days)	3	67
	4 (121-150 days)	4	33
3.25	1 (31-60 days)	1	60
	2 (61-90 days)	2	50
	3 (91-120 days)	3	50
	4 (121-150 days)	4	50

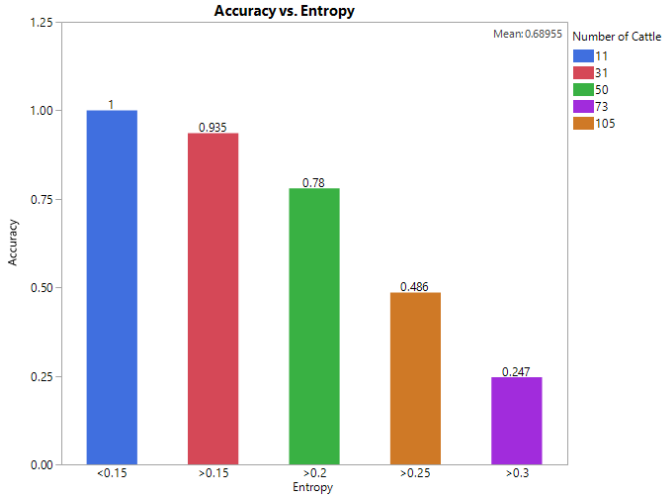


Fig. 7. The estimation accuracy of proposed BNM with higher confidence

According to entropy calculation, when the proposed model gets higher confidence (0.15<entropy>0.15), the accuracy rate lies in between 93% to 100%, which is one of the most significant finding in this research. Fig. 7 represents the results of the proposed BNM with higher confidence. In addition, Fig. 8 illustrates briefly the most influential parameters distribution according to the accuracy estimation with high confident model. These distributions would assist the farm management to find out most optimum cattle to have estrous cycle on due course.

All of these findings represent significant reliability and confidence to estimate the presence of estrous cycle with comparison to other traditional methods. In addition, all these sample cattle were previously anestrus due to some health hazards and other reasons in their farm. This investigation method could easily deploy to other healthy cattle for designing proper estrous synchronization protocol too. Moreover, the analytical approach of using Bayesian network discovers the most optimum conditions for each individual

cattle to find the presence of estrous cycle. The objective estimation of finding out the presence or absence of estrous cycle in cattle definitely boosts up the productivity in this arena and as well as leads to a new field of research.

V. Conclusion and Future Plan

This research presented the discovering method of cattle's estrous cycle presence using a new approach of Bayesian network model. The inclusion of body condition parameters, postpartum intervals and parity in the model helped to evaluate more accurate objective estimation of estrous cycle presence in cattle. The results and analysis confirmed that, the more accurate and optimum factors for cattle productivity could be found by the proposed BNM. The authors believe, the objective estimation definitely provides boost-up in the productivity of livestock industry in Japan and others countries. In future, the authors would like to include other parameters of cattle for finding out the presence of estrous cycle. In addition, the proposed model could be validated by using more sample datasets in future. Therefore, the proposed methods would get higher confidence and reliability to use at industry level.

Acknowledgment

This investigation is funded and supported by Ministry of Agriculture, Forestry and Fisheries (MAFF), Japan. The authors also show gratitude and appreciation to the Morinaga Dairy Service Co. Ltd., Japan for their constant support during this investigation.

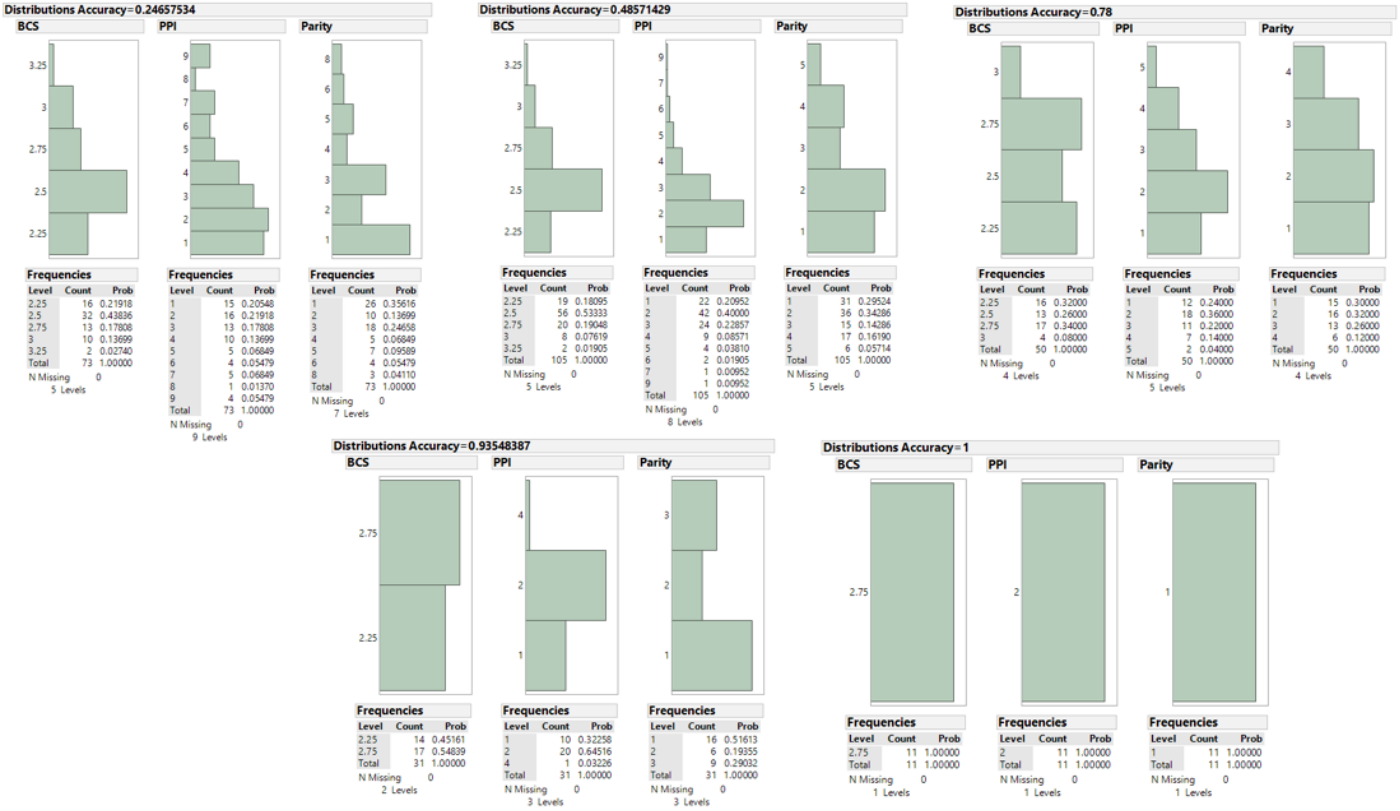


Fig. 8. Distribution of BCS, PPI and Parity according to high confident model for estimating the presence or absence of estrous cycle

References

[1] J. C. Whittler, "Reproductive Anatomy and Physiology of the Cow", Department of Animal Sciences. University of Missouri, Accessed December, 2015.

[2] P. D. Burns, "The Dairy Cow Heat Cycle", Colorado State University, Accessed December, 2015.

[3] J. A. Parish, J. E. Larson, and R. C. Vann, "The Estrous cycle of Cattle", Mississippi State University in cooperation with US Department of Agriculture, Publication No.2616, 2010.

[4] G. Perry, "The Bovine Estrous Cycle- FS921A", South Dakota State University-Cooperative Extensive Service USDA, Accessed December, 2015.

[5] J. Walker, and G. Perry, "Cow Condition and Reproductive Performance", Proceeding of The Range Beef Cow Symposium XX, Colorado, USA, December, 2007.

[6] L. F. M Pfeifer, S.C.B.S. Leal, A. Scheneider, E. Schemitt, and M.N. Correa, "Effect of ovulatory follicle diameter and progesterone concentration on the pregnancy rate of fixed time inseminated lactating beef cows", Revista Brasileria de Zootecnia, Vol. 41, No. 4, 2012, pp. 1004-1008.

[7] M. Matsui, and A. Miyamoto, "Evaluation of ovarian blood flow by colour Doppler Ultrasound: Practical use for reproductive management in the cow", The Veterinary Journal, 181, 2009, pp.232-240.

[8] T.A.Zacarias, S.B. Sena-Natto, A.S. Mendonca, M.M. Franco, and R.A. Figueiredo, "Ovarian Follicular Dynamics in 2 to 3 months old Nelore Calves (Bos Taurus indices)", Journal of Animal Reproduction, Vol. 12, No.2, June,2015, pp.305-311.

[9] G.A. Perry, M.F. Smith, A.J. Roberts, M.D. MacNeil, and T.W. Geary, "Relationship between size of the ovulatory follicle and pregnancy success in beef heifers", Journal of Animal Science, 85:684-689, 2007.

[10] A. Honnens, C. Voss, K. Herzog, H. Niemann, D. Rath, and H. Bollwein, "Uterine blood flow during the first 3 weeks of pregnancy in Dairy Cows" Journal of Theriogenology, Vol.70, 2008. Pp.1048-1056.

[11] G. Campanile, G. Neglia, R. Di Palo, B. Gasparrini, C. Pacelli, M. D'Occhio, and L. Zicarelli, "Relationship of body condition score and blood urea and ammonia to pregnancy in Italian Mediterranean buffaloes", Reproduction Nutrition Development, EDP Sciences, 2006, 46 (1), pp.57-62.

[12] G. A. Perry, O. L. Swanson, E. L. Larimore, B. L. Perry, G. D. Djira, and R. A. Cushman, "Relationship of follicle size and concentrations of estradiol among cows exhibiting or not exhibiting estrus during a fixed-time AI protocol", Journal of Domestic Animal Endocrinology, 48(2014), pp.15-20.

[13] W. Kellogg, "Body Condition Scoring with dairy cattle- FAS4008", University of Arkansas, USA, Accessed on: January 2016.

[14] J.M. Bewley, and M.M. Schutz, "Review: An interdisciplinary review of Body Condition Scoring for Dairy Cattle", The Professional Animal Scientist 24(2008), pp. 507-529.

[15] F.C. Castro, J.O. Porcayo, R.J. Ake-Lopez, J.G.M. Monforte, R.C. Montes-Perez, and J.C.S. Correa, "Effect of Body Condition Score on Estrous and Ovarian function characteristics of Synchronized Beef-Master Cows", Journal of Tropical and Subtropical Agroecosystems, 16(2013), pp.193-199.

[16] K. Yamada, T. Nakao, and N. Isobe, "Effects of Body Condition Score in Cows Peripartum on the onset of the Postpartum Ovarian Cyclicity and Conception rates after Ovulation Synchronized/ Fixed-Time Artificial Insemination", Journal of Reproduction and Development, Vol. 49, No. 5, 2003, pp.381-388.

[17] M. DeJarnette, "Estrus Synchronization: A Reproductive Management Tool", White Paper, Select Sires Inc., Ohio, USA, 2004.

[18] M. Takagi, N. Yamagishi, I.H. Lee, K. Oboshi, M. Tsuno, and M.P.B. Wijayagunawardane, "Reproductive management with Ultrasound Scanner Monitoring System for a high-yielding Commercial Diary Herd Reared under Stanchion Management Style", Asian-Australian Journal of Animal Science, 2005, Vol. 18, No. 7, pp. 949-956.

[19] G.P. Adams, and J. Singh, "Bovine Bodyworks: ultrasound Imaging of Reproductive Events in Cows", WCDS Advances in Dairy Technology, Vol. 23, 2011, pp. 239-254.

[20] J.H.M. Viana, E.K.N. Arashiro, L.G.B. Siqueira, A.M. Ghetti, V.S. Areas, C.R.B. Guimaraes, M.P. Palhao, L.S.A. Camargo, and C.A.C Fernandes, "Doppler Ultrasonography as a tool for Ovarian Management", Journal of Animal Reproduction, Vol. 10, No. 3, September 2013, pp. 215-222.

[21] P.M. Fricke, and G.C. Lamb, "Practical applications of ultrasound for reproductive management of beef and diary cattle", Proceedings of The Applied Reproductive Strategies in Beef Cattle Workshop, Kansas, USA, September 2002.

[22] G.C. Lamb, C.R. Dahlen, and D.R. Brown, "Reproductive Ultrasonography for monitoring Ovarian Structure Development, Fetal Development, Embryo Survival and Twins in Beef Cows", The Professional Animal Scientist Symposium, No. 19, 2003, pp. 135-143.

[23] Anonymous, "Body Condition Scoring in Dairy Cattle- AI10782", White Paper, Elanco Animal Health, 1-800-428-4441, 2009.

[24] J.D. Ferguson, D.T. Galligan, and N. Thousen, "Principal Descriptor of Body Condition Score in Holstein Cows", Journal of Dairy Science, No.77, 1994, pp.2695-2703.

[25] Report of National Livestock Breeding Center, Japan. Website: http://www.nlbc.go.jp/en/, Accessed January, 2016.

[26] W.J. Burkholder, "Use of Body Condition Scores in Clinical Assessment of the Provision of the Optimal Nutrition", JAVMA, Vol.217, No. 5, September 2000.

[27] O.Fukuda, T.Tsuji, and M.Kaneko, "A human supporting manipulator based on manual control using EMG signal", Journal of the Robotics Society, Japan, Vol.18, No.3, 2000, pp.79-86.

AUTHORS PROFILE

Kohei Arai, He received BS, MS and PhD degrees in 1972, 1974 and 1982, respectively. He was with The Institute for Industrial Science and Technology of the University of Tokyo from April 1974 to December 1978 also was with National Space Development Agency of Japan from January, 1979 to March, 1990. During from 1985 to 1987, he was with Canada Centre for Remote Sensing as a Post Doctoral Fellow of National Science and Engineering Research Council of Canada. He moved to Saga University as a Professor in Department of Information Science on April 1990. He was a councilor for the Aeronautics and Space related to the Technology Committee of the Ministry of Science and Technology during from 1998 to 2000. He was a councilor of Saga University for 2002 and 2003. He also was an executive councilor for the Remote Sensing Society of Japan for 2003 to 2005. He is an Adjunct Professor of University of Arizona, USA since 1998. He also is Vice Chairman of the Commission "A" of ICSU/COSPAR since 2008. He wrote 33 books and published 510 journal papers. He is now Editor-in-Chief of IJACSA and IJISA.

Osamu Fukuda received his B.E. degree in mechanical engineering from Kyushu Institute of Technology, Iizuka, Japan, in 1993 and the M.E. and Ph.D. degrees in information engineering from Hiroshima University, Japan in 1997 and 2000, respectively. From 1997 to 1999, he was a Research Fellow of the Japan Society for the Promotion of Science. He joined Mechanical Engineering Laboratory, Agency of Industrial Science and Technology, Ministry of International Trade and Industry, Japan, in 2000. Then, he was a member of National Institute of Advanced Industrial Science and Technology, Japan from 2001 to 2013. Since 2014, he has been a Professor of Graduate School of Science and Engineering at Saga University, Japan. Prof. Fukuda won the K. S. Fu Memorial Best Transactions Paper Award of the IEEE Robotics and Automation Society in 2003. His main research interests are in human interface and neural networks. Also, he is currently a guest researcher of National Institute of Advanced Industrial Science and Technology, Japan. Prof. Fukuda is a member of IEEE and the Society of Instrument and Control Engineers in Japan.

Iqbal Ahmed got his Bachelor of Science (BSc) Honors degree in Computer Science and Engineering from University of Chittagong, Bangladesh in 2007 and achieved joint Master degree from PERCCOM program of European Union in September 2015. He received his Master of Complex System Engineering degree from University of Lorraine (UL), France then Master in Technology from Lappeenranta University of Technology (LUT), Finland and Master degree in Pervasive Computing and Communication for Sustainable development from Lulea University of Technology (LTU), Sweden. Since October 2015, he is enrolled as a doctoral student in the Department of Information Science, Saga University, Japan. In profession, he worked in the Department of Computer Science and Engineering, University of Chittagong, Bangladesh as an Assistant professor since February 2011. He has been awarded Cat-A scholarship of Erasmus Mundus from European Union two times in 2010 and 2013 respectively. His current research interest lies in the field of green and sustainable computing and information processing.

Analysis of Security Requirements Engineering: Towards a Comprehensive Approach

Ilham Maskani[1]
LISER Laboratory
ENSEM, Hassan II University
Casablanca, Morocco

Jaouad Boutahar[2], Souhaïl El Ghazi El Houssaïni[3]
Systems, architectures and networks Team
EHTP
Casablanca, Morocco

***Abstract*—Software's security depends greatly on how a system was designed, so it's very important to capture security requirements at the requirements engineering phase. Previous research proposes different approaches, but each is looking at the same problem from a different perspective such as the user, the threat, or the goal perspective. This creates huge gaps between them in terms of the used terminology and the steps followed to obtain security requirements. This research aims to define an approach as comprehensive as possible, incorporating the strengths and best practices found in existing approaches, and filling the gaps between them. To achieve that, relevant literature reviews were studied and primary approaches were compared to find their common and divergent traits. To guarantee comprehensiveness, a documented comparison process was followed. The outline of our approach was derived from this comparison. As a result, it reconciles different perspectives to security requirements engineering by including: the identification of stakeholders, assets and goals, and tracing them later to the elicited requirements, performing risk assessment in conformity with standards and performing requirements validation. It also includes the use of modeling artifacts to describe threats, risks or requirements, and defines a common terminology.**

Keywords*—Security requirements; Requirements engineering; Security standards; Comparison; Risk assessment

I. Introduction

Security needs have evolved with the evolution of information systems (IS). IS are more and more open and interconnected, which makes securing these IS more necessary and more challenging. But, in the Software Development Life Cycle (SDLC), security issues are often addressed at the design phase at best, or at maintenance phase at worst by fixing detected vulnerabilities. As reported in this paper [1], finding and fixing a software problem after delivery is often 100 times more expensive than finding and fixing it during the requirements and design phase. A model developed by MIT, whose objective is to prove the return of investment on secure software development, showed that the earliest the security is addressed, the highest the benefit (21%)[2]. Thus, it is critical to address security issues at the earliest phase. This is the reason why OWASP recommends focusing a big part of security flaws detecting efforts on the requirements engineering phase and the design phase as shown in fig. 1[3]. Requirements engineering is the very first step to make any software. It is usually applied to functional requirements, and can be extended to quality and security requirements, traditionally considered non-functional. By integrating security requirements into requirements engineering, a big improvement can be made in term of security vulnerabilities, software maintenance efforts and development costs. Many initiatives propose different approaches to security requirements engineering (SRE), along with literature reviews of these approaches. In the first section, these works will be presented. The term "approach" will be used to refer to any method, framework, etc. which sets out clear steps to obtain security requirements. In the second section, the selection and comparison process followed for featured SRE approached will be explained. Then, approaches will be compared according to the predefined criteria. In the final section, a common terminology will be defined for the concepts used by the approaches. Then the outline of our comprehensive approach to SRE will be presented, along with its desired qualities.

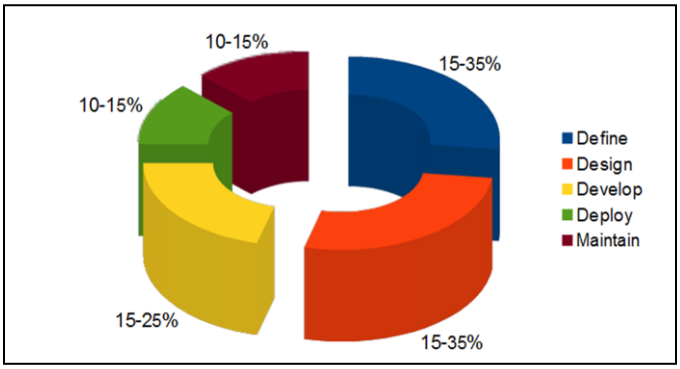


Fig. 1. Recommended proportions of Test Effort in SDLC

II. Related Work

To achieve our aim, relevant literature reviews were studied and primary approaches were compared to find their common and divergent traits. This section presents the reviews and approaches featured in our research. These approaches were selected by applying the selection & comparison process detailed in the next section.

A. Reviews

1) Survey and analysis on Security Requirements Engineering: [4] It is the most recent detailed analysis on the subject. They discuss various types of security requirements with given examples, stretching the importance of considering security requirements as functional requirements. They compare approach activities to identify the weaknesses of

each. The choice of an approach over another depends on covered activities and existing SW development methods in an organization.

2) A Comparison of SRE methods: [5] proposed a conceptual framework against which approaches can be evaluated. They made a commendable effort to categorize existing approaches: Multi-view approaches, Goal-based approaches …

3) A systematic review of security requirements engineering : [6]A systematic, thorough review which aims to supply researchers with a summary of all the existing information about security requirements in a thorough and unbiased manner, providing a background in which to appropriately position new research.

4) Security Requirements for the Rest of Us - A Survey : [7]This survey highlights mainstream approaches. It focuses on the importance of simplifying SRE methods since a lightweight method is more likely to be adopted than a complex one. It also stretches the importance of scholar education of developers and software engineers on the SRE discipline.

B. Overview of Approaches

1) SREF : Security Requirements Framework by Haley et al. [8] is a mix between engineering requirements and security requirements. It's iterative as it goes back and forth between modeling and requirements engineering. SREF follows 4 steps:

- Identify functional requirements
- Identify security goals
    - Identify assets
    - Generate threat description
    - Apply management principles (separation of duties, functions, ..)
- Identify security requirements: constraints on one or more security goal. The security requirements are denoted textually.
- Construct satisfaction arguments: show that the system can satisfy the security requirements.

2) KAOS anti-models: To elaborate security requirements, Van Lamsweerde suggests using KAOS by constructing intentional anti-models. KAOS is a Goal Oriented Method for requirements engineering. A goal is a desired property of the IS to be, that has been expressed by a stakeholder. The satisfaction of this goal will depend on successful cooperation between all agents of the IS. KAOS documents requirements using a goal tree, with strategic goals as the root and IS requirements as leafs. Security requirements using anti-models are elaborated in 3 steps. First, model the security goals. Then, derive from the former model an anti-model based on threats. Finally, derive from both former models countermeasures and define the security requirements. A requirement is defined as a terminal goal under the responsibility of an agent in the software.

3) MOSRE: The aim of the Model Oriented Security Requirements Engineering approach [9] is the use of models (App's use cases, misuse cases, …) to make the traceability and analysis of requirements easier. It's tailored for web applications. The particularity to MOSRE is that it encompasses identification of goals for the whole IS, the elicitation and the modeling of non-security requirements (functional or non-functional) before dealing with the security requirements. It is thus a method that can be applied to the whole requirements engineering phase, with a special focus on security. MOSRE steps are:

- Inception: Identify web app objectives, stakeholders and assets
- Elicitation
    - Elicit security and non-security goals and requirements
    - Identify threats and vulnerabilities
    - Risk assessment
    - Identify Security requirements
    - Generate Use case diagrams considering security requirements
- Elaboration : Generate structural analysis models (ex : data model, flow models) and develop UML diagrams to give a view of the secure web application in general (ex: high level class diagram, sequence diagram)
- Negotiation and validation of requirements

4) MSRA: The focus of the MSRA (Multilateral security requirements analysis) approach is to identify and analyze security requirements from the multiple views of stakeholders[10]. Security requirements result from the reconciliation of multilateral security goals, which are selected from a rich taxonomy. Security goals, and later requirements, contain the attributes "stakeholders" who have an interest in the requirement, "counter-stakeholders" towards whom a requirement is stated, and other attributes such as "owner", "degree of agreement" between stakeholders, the "information" to be protected by the requirement, the security "goal" that the requirement achieves… A singularity of MSRA is that, when resolving conflicts between requirements, it takes into account both functional (assumed to be extracted prior to applying MSRA) and security goals. There is a variant of MSRA, the Confidentiality Requirements Elicitation and Engineering (CREE) approach, which focuses only on confidentiality requirements and how they can be formalized. The steps followed by the MSRA are:

- Identify stakeholders
- Identify episodes: Episodes are similar to scenarios, but are of a lower granularity. They are used to partition the security goals and are later useful in identifying conflicts between multiple security goals.

- Elaborate security goals: Identify and describe the security goals of the different stakeholders for each of the episodes.
- Identify facts and assumptions: These are the properties of the environment that are relevant for stating security goals.
- Refine stakeholder views on episodes: Elaborate the stakeholder views taking facts, assumptions, and the relationships between episodes into account.
- Reconcile security goals: Identify conflicts between security goals, find compromises between conflicting goals, and establish a consistent set of security system requirements.
- Reconcile security and functional requirements: Trade functionality for security and vice versa in case of conflicting functional and security requirements.

5) Secure TROPOS: Tropos is a requirements-driven software development methodology. It's based on the i* framework, an agent-oriented modeling framework. While Tropos guides the development of agent-based systems through all phases of the SDLC, it is very focused on the requirements engineering phase. Secure Tropos[11] is based on the concepts of social relationships for defining the obligations of actors to other actors : functional dependency, ownership, provisioning, trust, and delegation of permission.
Secure Tropos steps are:

- Early requirements phase: studies the organizational setting of the future system
    - Actor diagram : identifying stakeholders and trust relationships between them (Trust modeling, Functional Modeling and Trust Management implementation)
    - Goal diagrams for each actor
- Late requirements phase: describes the future system within its operational environment, along with relevant functions and qualities, using further actor and goal diagrams.
- Requirements analysis :
    - Expressing system requirements in form of actors' properties and relations
    - validation of both functional and security requirements

Secure Tropos has been applied to the Italian data protection legislation compliance[12].

6) Holistic security requirements engineering: Holistic security requirements engineering [13] was conceived to overcome the shortcomings of other approaches to SRE. This approach, aimed at electronic commerce systems, defines risks, business processes and stakeholder & environmental demands as sources of security requirements. This leads to holistic security requirements, defined as "a need or restriction from a user, a stakeholder or the environment related to the goal to improve the system security".

The approach is described by this biphasic process with the following activities:

- Phase I: Preparation, aims to gather requirements from each of the sources.
    - Definition of goals
    - Security enhanced business modeling: Modeling business information exchange, considering security as business functionality.
    - Requirement transformation: transforming security considerations from the business model into security requirements
    - Internal requirement elicitation: Detailing the transformed requirements
    - Stakeholder definition
    - Requirement elicitation: From stakeholders' points of view
    - Risk assessment : through a baseline investigation of risks using checklists
- Phase II: Compilation, aims to compile the different requirements and resolve conflicts between them.
    - Compilation
    - Formal security requirements specification
    - Prototyping
    - Validation

An evolution of this approach, named SKYDD, was developed to better suit the needs of telecom providers.

7) SQUARE: Developed by Carnegie Mellon University, SQUARE (Security Quality Requirements Engineering)[14] is a 9-steps process whose goal is to get categorized and prioritized security requirements.

Each step is described with inputs, outputs, participants and techniques:

- Agree on definitions
- Identify security goals
- Develop Artifacts to support security requirements definition
- Perform risk assessment
- Select elicitation techniques
- Elicit security requirements
- Categorize requirements
- Prioritize requirements
- Requirements inspection

This approach had been extended to specifically treat privacy (P-SQUARE) and acquisition (A-SQUARE).

8) SREP: Security Requirements Engineering Process[15] is a process centered on the security evaluation standard Common Criteria[16] and based on the notion of reuse. It deals with security requirements in a systematic and intuitive way. It provides a security resources repository and integrates the Common Criteria into the software lifecycle, so that it unifies the concepts of requirements engineering and security engineering. In order to support this approach, many concepts and techniques are used: a security resources repository (with

assets, threats, requirements, etc), misuse cases, threat/attack trees, and security uses cases. SREP has been developed by taking into account the standard ISO/IEC 27002[17].

SREP activities are:

- Agree on Definitions
- Identify Vulnerable &/or Critical Assets
- Identify Security Objectives & Dependencies
- Identify Threats & Develop Artifacts
- Risk Assessment
- Elicit Security Requirements
- Categorize & Prioritize Requirements
- Requirement Inspection
- Repository Improvement

SREPPLINE is a declination of SREP specific to Software Product Lines.

9) STS: Going from the statement that software operates within the context of larger socio-technical systems, STS is an approach for modeling and reasoning about security requirements in such systems [18]. Security requirements are specified, via the STS-ml requirements modeling language, as contracts that constrain the interactions among the actors. The requirements models of STS-ml have a formal semantics which enables automated reasoning for detecting possible conflicts among security requirements. STS was applied to an e- Government system for tax collection.

STS steps are:

- Model system components and interaction with STS-ml language
    - Social view for stakeholders
    - Information view
    - Authorizations view
- Use the models to specify security requirements as constraints on the interactions. Security requirements are specified in the STS-ml language.
- Use the automated reasoning to detect conflicts

III. Comparison of SRE Approaches

This section presents the process followed to select and compare the approaches featured in our research and shows the results of the comparison.

A. Comparison Process

To guarantee the comprehensiveness of our approach, a documented selection and comparison process was followed (see fig. 2). This process is inspired by an evaluation method for engineering approaches in the secure SDLC named SecEval [19]. This distinguishes our work from the previous reviews as they compare only a certain set of approaches, without explaining the inclusion or exclusion criteria. Documenting our process makes this comparison reproducible for future research.

1) Sources : The aforementioned reviews were a very rich source. To complete the information gathered, we queried different scientific databases to find novel research in the area. This way, we obtained other approaches that have not yet been featured in any of the previous reviews, such as MOSRE and STS. Other sources were: Sciencedirect, ResearchGate and GoogleScholar.

2) Selection criteria: Selection criteria were applied on the gathered research. The first criterion is if the proposed approach is focused on the early phase of the development lifecycle. Indeed, many approaches go straight to the design phase by proposing modeling approaches, without specifying how to extract those requirements in the first place. Others propose activities to enhance security through the whole Software Development Life Cycle such as CLASP[20] and Microsoft SDL [21]. To have a precise scope, only the methods that focus on the requirements engineering phase were kept. The second criterion is the novelty. Chosen approaches have been referenced in the years 2008 and up. The third criterion is that chosen approaches offer a clear process or clear steps about how to extract the security requirements, and not just general guidelines about security requirements, or their management.

3) Information extraction: Once the final approaches were selected, the following information was extracted to be used as comparison criteria.

- Steps: What are the clear steps followed to obtain security requirements
- Security Objectives: Whether the approach addresses all security objectives (Confidentiality, Integrity, Availability, …) or focuses on a single one
- Tool / Notation support: Whether there is a tool or a notation developed to support the use of the approach
- Use / Application: Whether the approach had been applied to a case study or a real IS.

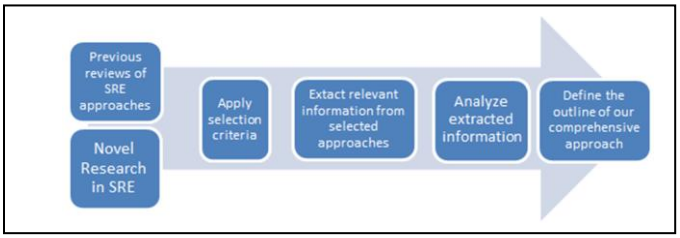


Fig. 2. Selection & Comparison process

- Includes modeling activities (design): Whether the approach includes high level design activities, taking into account the obtained security requirements.
- Compliance with security standards: Whether the approach is compliant or inspired by any security standard
- Reusability of requirements: Whether the approach promotes the reuse of obtained requirements

- Use of ontology/taxonomy: Whether the approach uses an existing ontology ot taxonomy to define the approach steps and to define the security requirements
- Domain specific: Whether the approach is dedicated to a certain type of software (Web applications, Mobile, E-Gov, etc.)

B. Comparison Results:

Tab. 1 summarizes the steps found in each approach, and gives a synthetic view about the most and the least common steps. No single approach includes these steps all at once. First, we can see that "Identifying vulnerabilities/threats" and "Identifying security goals" are the most common steps since we can't derive requirements without establishing goals, and it's important to know a system's vulnerabilities and threats to be able to secure it. Then, other steps are also quite persistent such as "Identifying stakeholders", "Creating security artifacts" and "Validation of requirements". Identifying stakeholders is a way to make sure that all the systems goals have been mapped, since different stakeholders will have different views of the systems, and thus different goals. Creating security artifacts is important as it helps clarify the requirements by incorporating artifacts such as attack trees and misuse cases. It will also help designers and developers during later phases of the project. As for Requirements validation, the goal of it is to make sure that all goals have been covered by the elicited requirements, with no conflicts between them. Finally, some steps are often neglected even if they're very important, such as "Risk assessment" and "Repository enhancement". Risk assessment builds on the identified threats and vulnerabilities to identify analyze and evaluate risks by choosing for example the risks to accept and those to mitigate. Assessing risks leads to thinking about security controls, which could lead to new requirements. Keeping and enhancing a repository is a way to promote de reuse of requirements. Such a repository can be used to validate the obtained requirements and identify new ones.

As for the characteristics comparison, we present in tab. 2 the results for each approach regarding to the aforementioned comparison criteria. First thing we deduce is that there is no approach that fulfills all criteria. Apart from Secure Tropos, all approaches try to cover most security objectives, especially the CIA triad (Confidentiality, Integrity, and Availability). Some approaches are defined from the beginning to better suit certain systems such as Web Applications that are more and more used to replace custom applications. When applied, they are aimed at highly data sensitive systems such as e-gov, e-commerce and e-health. As for artifacts and notation, the most used are UML based (misuse cases, UMLSec [22]) and attack trees. Some approaches have developed their own notation system, or even a tool to create their artifacts and support their approach. The conformity to security standards is quite present, especially for the approaches that include risk assessment. Common security standards used are the ISO 27000 family of standards[23] and the SSE-CMM (Systems Security Engineering- Capability Maturity Model)[24]. For the purpose of better understanding of requirements, some approaches propose their own format in which requirements are documented. The rarest characteristics were the use of a taxonomy or ontology to build the approach, and the existence of a tool supporting the approach.

IV. Outline of Our Comprehensive Approach

A. Common Terminology:

From studying each approach, we can identify a set of concepts that are consistent through most approaches: Stakeholder, Asset, Risk, etc… These concepts are drawn from both the fields of security and requirements engineering. Tab. 3 below offers a definition of these concepts to establish a common terminology based on the ISO/IEC 27000:2016 vocabulary[25]. Some existing papers offer detailed taxonomies [26]and facilitate applying SRE approaches. This is the terminology that we will base our approach on.

B. Proposed Activities

Based on the previous section, we can give guidelines about a new comprehensive approach that takes into account the strengths and weaknesses of studied approaches. We will try to avoid being too specific about a domain or any other specificity that might limit the use of our approach. Still, the new approach has to include important concepts and techniques such as: identification of stakeholders, identification of assets and threats, risk assessment and reuse of requirements. It also has to follow general guidelines of requirements engineering by documenting, tracing and validating requirements. These are the activities that we propose for our approach:

1) Identify stakeholders
2) Identify assets
3) Identify Security goals
4) Identify Threats/vulnerabilities
5) Create artifacts: Misuse cases, attack trees, etc.
6) Risk assessment (in conformity to ISO/IEC 27005)
7) Elicit security requirements Format security requirements
8) Categorize and Prioritize
9) Inspection/validation
*10)*Enhance IS Use case by including security (ex : UML sec)
*11)*Repository Enhancement

TABLE I. OCCURENCES OF STEPS

Steps	Approaches									
	SREF	**KAOS anti-models**	**MOSRE WebApp**	**MSRA**	**Secure Tropos**	**Holistic SRE**	**SQUARE**	**SREP**	**STS**	**Number of occurrences**
Agree on definitions							X	X		2/9
Identify assets	X	X	X					X		4/9
identify stakeholders		X	X	X	X	X			X	6/9
Identify security goals/objectives	X	X	X	X	X	X	X	X	X	9/9
identify business/ IS objectives	X		X			X				3/9
Identify threats	X	X	X		X		X	X	X	7/9
Develop Artifacts		X	X		X		X	X	X	6/9
Perform risk assessment			X			X	X	X		4/9
Select elicitation techniques			X				X			2/9
Elicit -non security requirements	X		X							2/9
Elicit security requirements	X	X	X	X	X	X	X	X	X	9/9
Categorize / Prioritize requirements			X				X	X	X	4/9
Requirements inspection/validation/Conflict resolution	X		X	X		X	X	X	X	7/9
Repository Improvement								X		1/9

TABLE II. CHARACTERISTICS OF APPROACHES (COMPARISON CRITERIA)

COMPARISON CRITERIA	APPROACHES								
	Holistic SRE	**KAOS anti-models**	**MOSRE WebApp**	**MSRA**	**Secure TROPOS**	**SREF**	**SREP**	**SQUARE**	**STS**
Security Objectives Specific	confidentiality, integrity, non-repudiation	CIA + privacy, authentication, non-repudiation		CIA + accoutability, pseudonimity	Privacy, Trust				CIA + accountability, reliability, authenticity
Tool / Notation support	No	Temporal logic notations	No	No	Si*, ST-tool	No		P-square	STS-ml, STS Tool
Use / Application	e-Commerce, Telecom	e-Banking	e-Voting, e-Healh system	e-Health	Italian Legislation compliance	No	Software Product Lines	Asset Management System	e-Government
Includes modeling activities of requirements	Yes	Yes	Security use cases, misuse cases, attack trees		Yes	No	Security use cases, misuse cases, attack trees	misuse cases, attack trees	Yes
Compliance with security standards	ISOO 27000, SSE-CMM	No		No	ISO/IEC 27002	No	Common Criteria, SSE-CMM, ISO/IEC 27002	NIST SP 800-30	ISO 27005
Format / Reusability of requirements	Yes	No	Yes	Yes	No	No	Yes	No	Yes
Based on ontology or taxonomy	No	No	No	Yes	No	No		No	
Domain specific	e-Commerce, Telecom	No	Web Apps	No	Agent based systems	No		No	Large socio-technical systems

If those activities are followed correctly, our approach would have the following qualities:

- Environment reconnaissance: The more complex the IS, the more important it is to identify the stakeholders and the assets. Elicited security requirements will have to be traced all the way back to the related assets and related stakeholders.
- Risk assessment: The finality of securing a system is to be prepared against all risks. Thus, it is important for our approach to identify all vulnerabilities and threats, to enable a thorough risk assessment.

- Favor re-usable requirements :
    - Propose a standard format to represent security requirements.
    - Keep a repository of sample and categorized requirements
- Follow the fundamentals of requirements engineering. Some of those fundamentals tend to be overlooked:
    - Traceability: It is important to be able to match each obtained requirement with the associated risk, the asset, the security goal it covers and the stakeholder who expressed it. This will help at the requirements inspection phase, and at later phases of the SDLC when managing requirements.
    - Inspection and validation: Obtained requirements should be inspected to resolveany conflicts, and to ensure complete coverage of all the initially stated security goals.

TABLE III. COMMON TERMINOLOGY

Concept	**Definition**	**Alternate labels**
Stakeholder	Person or organization that can affect, be affected by, or perceive themselves to be affected by a decision or activity. Some approaches include other systems that have an interest in the IS.	Actor, client, agent
Asset	Anything that has value to the organization, its business operations and their continuity, including Information resources that support the organization's mission (Data).	Information, Resource, Object
Goal	A Security objective that must be achieved by the system to be	Objective
Vulnerability	weakness of an asset or control that can be exploited by one or more threats	
Threat	potential cause of an unwanted incident, which may result in harm to a system or organization	
Risk	Potential that threats will exploit vulnerabilities of an information asset or group of information assets and thereby cause harm to an organization	
Risk Assessment	Overall process of risk identification, risk analysis and risk evaluation	Risk identification, risk analysis, risk evaluation
Requirement	Need or expectation that is stated, generally implied or obligatory. Requirements are low level details of goals.	Goal, objective
Control	Measure that is modifying risk	Countermeasure
Attack	Attempt against the security of an asset	

- Easy and faithful transition from requirements engineering phase to design phase: Use of modeling artifacts to describe threats, risks and requirements.
- Use of existing risk management standard and Bodies Of knowledge (ISO 27002, ISO 27005, EBIOS, BSI, etc.) for threats, risk assessment and security goals.
- Ease of use: It should be detailed and documented enough to be applied easily. Complicated and time consuming steps (ex: modeling artifacts) should be simplified and kept to a minimum.

V. CONCLUSION & PERSPECTIVES

Our aim was to define the outline of a comprehensive approach to security requirements engineering. To achieve that, a thorough analysis of existing SRE approaches was conducted. The outline, along with a common terminology, was drawn from this analysis. The first contribution of our research is that it can be used by fellow researchers or practitioners to position themselves between heterogeneous approaches. Our comparison criteria and common terminology allows a better understanding of each approach, and can help choose the most appropriate approach for a certain need. The second contribution is our comprehensive approach that conciliates between the different trends to security requirements engineering: goal oriented, risk analysis oriented and multilateral. As such, it distinguishes itself by being faithful to the fundamentals of requirements engineering, to security standards and by facilitating the use of security requirements in later phases of the SDLC through requirements formatting and security enhanced system artifacts. When eliciting requirements, regardless of the approach used, security requirements shouldn't be an afterthought, but an indivisible part of requirements engineering for the system as a whole. Security requirements should be confronted with other functional, quality or performance requirements for further validation and conflict resolution so they would be incorporated in the system's design.

Our plans for future work are to fully develop our approach following the described outline. We would document the inputs, activities and outputs of each step, describe the artifacts to be created, and develop a format for security requirements. We would also explain how our approach integrates with security in later phases of the SDLC. We plan to validate our approach by applying it to a concrete security sensitive system, and measure security metrics to improve its efficiency.

REFERENCES

[1] B. Boehm and V. R. Basili, "Top 10 list [software development]," Computer, vol. 34, no. 1, pp. 135–137, 2001.

[2] H. S. Venter and Information Security South Africa, Eds., Peer-reviewed proceedings of the ISSA 2004 enabling tomorrow conference. ISSA, 2004.

[3] "Testing Guide Introduction - OWASP." [Online]. Available: https://www.owasp.org/index.php/Testing_Guide_Introduction. [Accessed: 13-Oct-2016].

[4] P. Salini and S. Kanmani, "Survey and analysis on Security Requirements Engineering," Comput. Electr. Eng., vol. 38, no. 6, pp. 1785–1797, Nov. 2012.

[5] B. Fabian, S. Gürses, M. Heisel, T. Santen, and H. Schmidt, "A comparison of security requirements engineering methods," Requir. Eng., vol. 15, no. 1, pp. 7–40, Mar. 2010.

[6] D. Mellado, C. Blanco, L. E. Sánchez, and E. Fernández-Medina, "A systematic review of security requirements engineering," Comput. Stand. Interfaces, vol. 32, no. 4, pp. 153–165, Jun. 2010.

[7] A. Tondel, M. G. Jaatun, and P. H. Meland, "Security Requirements for the Rest of Us: A Survey," IEEE Softw., vol. 25, no. 1, pp. 20–27, Jan. 2008.

[8] C. B. Haley, R. Laney, J. D. Moffett, and B. Nuseibeh, "Security Requirements Engineering: A Framework for Representation and Analysis," IEEE Trans. Softw. Eng., vol. 34, no. 1, pp. 133–153, Jan. 2008.

[9] P.Salini and S. Kanmani, "Security Requirements Engineering Process for Web Applications," Procedia Eng., vol. 38, pp. 2799–2807, 2012.

[10] S. F. Gürses and T. Santen, "Contextualizing Security Goals: A Method for Multilateral Security Requirements Elicitation.," in ResearchGate, pp. 42–53, 2006.

[11] P. Giorgini, F. Massacci, J. Mylopoulos, and N. Zannone, "Requirements engineering for trust management: model, methodology, and reasoning," Int. J. Inf. Secur., vol. 5, no. 4, pp. 257–274, 2006.

[12] F. Massacci, M. Prest, and N. Zannone, "Using a security requirements engineering methodology in practice: The compliance with the Italian data protection legislation," Comput. Stand. Interfaces, vol. 27, no. 5, pp. 445–455, Jun. 2005.

[13] Zuccato, "Holistic security requirement engineering for electronic commerce," Comput. Secur., vol. 23, no. 1, pp. 63–76, Feb. 2004.

[14] Mead N, Hough E, Stehney T (2005) Security quality requirements engineering (SQUARE) methodology. Carnegie Mellon Software Engineering Institute, Technical report CMU/SEI-2005-TR-009.

[15] D. Mellado, E. Fernández-Medina, and M. Piattini, "A common criteria based security requirements engineering process for the development of secure information systems," Comput. Stand. Interfaces, vol. 29, no. 2, pp. 244–253, Feb. 2007.

[16] "ISO/IEC 15408-1:2009 - Information technology -- Security techniques -- Evaluation criteria for IT security -- Part 1: Introduction and general model," ISO. [Online]. Available: http://www.iso.org/iso/home/store/catalogue_tc/catalogue_detail.htm?csnumber=50341. [Accessed: 25-Oct-2016].

[17] "ISO/IEC 27002:2013 - Information technology -- Security techniques -- Code of practice for information security controls," ISO. [Online]. Available: http://www.iso.org/iso/home/store/catalogue_tc/catalogue_detail.htm?csnumber=54533. [Accessed: 20-Oct-2016].

[18] E. Paja, F. Dalpiaz, and P. Giorgini, "Modelling and reasoning about security requirements in socio-technical systems," Data Knowl. Eng., vol. 98, pp. 123–143, Jul. 2015.

[19] M. Heisel, W. Joosen, J. Lopez, and F. Martinelli, Eds., Engineering Secure Future Internet Services and Systems, vol. 8431. Cham: Springer International Publishing, 2014.

[20] "CLASP Concepts - OWASP." [Online]. Available: https://www.owasp.org/index.php/CLASP_Concepts. [Accessed: 20-Oct-2016].

[21] "Microsoft Security Development Lifecycle." [Online]. Available: https://www.microsoft.com/en-us/sdl/. [Accessed: 20-Oct-2016].

[22] J. Jrjens, Secure Systems Development with UML. Berlin, Heidelberg: Springer-Verlag, 2010.

[23] "ISO/IEC 27001 - Information security management," ISO. [Online]. Available: http://www.iso.org/iso/home/standards/management-standards/iso27001.htm. [Accessed: 20-Oct-2016].

[24] "ISO/IEC 21827:2008 - Information technology -- Security techniques -- Systems Security Engineering -- Capability Maturity Model® (SSE-CMM®)," ISO. [Online]. Available: http://www.iso.org/iso/catalogue_detail.htm?csnumber=44716. [Accessed: 20-Oct-2016].

[25] "ISO/IEC 27000:2016 - Information technology -- Security techniques -- Information security management systems -- Overview and vocabulary," ISO. [Online]. Available: http://www.iso.org/iso/home/store/catalogue_tc/catalogue_detail.htm?csnumber=66435. [Accessed: 20-Oct-2016].

[26] N. Rjaibi and L. B. A. Rabai, "Developing a Novel Holistic Taxonomy of Security Requirements," Procedia Comput. Sci., vol. 62, pp. 213–220, 2015.

Teachme, A Gesture Recognition System with Customization Feature

Hazem Qattous
Department of Computer Information Systems
Applied Science Private University
Amman, Jordan

Bilal Sowan
Department of Computer Network Systems
Applied Science Private University
Amman, Jordan

Omar AlSheikSalem
Department of Software Engineering
Applied Science Private University
Amman, Jordan

***Abstract*—Many presentation these days are done with the help of a presentation tool. Lecturers at Universities and researchers in conferences use such tools to order the flow of the presentation and to help audiences follow the presentation points. Presenters control the presentation tools using mouse and keyboard which keep the presenters always beside the computer machine to be close enough to the keyboard and mouse. This reduces the ability of the lecturer to move close to the audiences and reduces the eye contact with them. Moreover, using such traditional techniques in controlling presentation tools lack the communication naturalness. Several gesture recognition tools are introduced as solutions for these problems. However, these tools require the user to learn specific gestures to control the presentation and/or the mouse. These specific gestures can be considered as a gestures vocabulary for the gesture recognition system. This paper introduces a gesture recognition system, TeachMe, which controls Microsoft PowerPoint presentation tool and the mouse pointer. TeachMe also has a gesture customization feature that allows the user to customize some gestures according to his/her preference. TeachMe uses Kinect device as an interface for capturing gestures. This paper, specifically, discusses in details the techniques and factors taken into consideration for implementing the system and its customization feature.**

Keywords*—Microsoft Kinect®; Gesture recognition system; Gesture customization

I. INTRODUCTION

These days, presentation tools, such as MS PowerPoint and Prezi, are used by lecturers to present lectures in a controlled flow. Presentation tools are also used in the domain of conferences [1]. Such presentations are controlled using keyboard and mouse which requires the lecturer to be standing beside the computer table all the time to start a presentation, move the presentation slides and ending the presentation. However, the lecturer still need to move back and forth between the computer and projection plan from one side and the students or the audiences for better communication from another side [2] and [3]. Additionally, using keyboard and mouse reduces the interactive and naturalness in communication with audiences [4] and [5]. References [2] and [1] introduce that even the Bluetooth connected devices, which are designed to control presentations have some drawbacks regarding the problem of keeping the lecturer close the computer because of its limited functionalities. As an example, such devices cannot control the mouse. A subjectively noticed problem of wireless connected devices is that the lecturer always forgets bringing the device to lecture. This returns the lecturer back to use keyboard and mouse.

All the above researchers introduced gesture capture devices as a solution for the above addressed problems. Although such gesture recognition devices were originally designed for interactive video games, several research found them suitable to be used as presentation controlling tools. Their suitability comes from their ability to provide an intuitive, natural and effective way for communication between human and machine [6] and [4]. Between all the research conducted in this field, very little focused on the problem of customizing the gesture. Basically, in any gesture recognition system, the implementation depends on a gesture vocabulary that is hard coded as that introduced by [7]. The user then should learn how to use the system based on the implemented vocabulary [8]. However, very little research focused on the customizing the gesture vocabulary to meet the preference of the user. This research addresses the customization problem and provide a simple solution that can be considered as an entrance for more research and implementations in this field. Providing the ability to customize gesture to a specific command is considered as the main contribution of this paper. This feature was implemented mainly for controlling MS PowerPoint presentation.

This paper presents a gesture recognition system, TeachMe. TeachMe is prototype that depends on Microsoft Kinect to recognize gestures. It is implemented to control Microsoft PowerPoint presentations and mouse pointer. The paper discusses implementation issues and obstacles faced during the system implementation. Additionally, it shows the parameters and factors taken into consideration during implementation. More importantly, the paper presents the new gesture customization feature implemented in TeachMe.

Next section of this paper presents the general structure and some general implementation information about TeachMe gesture recognition system. Implementation of TeachMe for controlling PowerPoint presentation and mouse pointer are then discussed in details. Later on, the paper presents the gesture customization feature implementation with some examples. After that, the paper introduces some related work and trials for implementing customization features at other gesture recognition systems. The limitations of TeachMe and its gesture customization features are discussed in detail

followed by a list the future work and plans for further development of the customization feature. Finally, the paper concludes and summarizes the work presented in the paper.

II. TEACHME GESTURE RECOGNITION SYSTEM

TeachMe is a research gesture recognition prototype that has been initially developed as a graduation project at Applied Science University. The system captures the lecturer body gestures through Microsoft Kinect® device. TeachMe development uses the standard Kinect SDK 1.8 as library. The Microsoft SDK uses three streams of data, RGB, depth data stream and skeleton data stream. As the coloring does not give useful information for this research purposes, it was neglected during implementation. The interest was directed to the depth and skeleton data streams. Once a gesture is captured, the system analyzes it. Analysis process maps the captured gestures to actions that are executed by the controlled part (presentation or mouse pointer). The mapping depends on a predefined gesture vocabulary that determines what gesture does which action. TeachMe recognizes two types of hand gestures. The first controls Microsoft PowerPoint presentation while the second controls the mouse pointer. Gestures controlling Microsoft PowerPoint includes starting a slideshow, moving slides forward and backward, and ending the slideshow. Gestures controlling mouse include moving mouse over the desktop, selecting an object (file or folder) by clicking on it, dragging an object, and double click on an object to open a file or explore a folder. Additionally, TeachMe provides the ability to customize some gestures to suit a lecturer preference. Customization feature will update the mapping between the recognized gesture and the action to be executed in the gesture analysis process.

III. CONTROLLING MS POWERPOINT

The system recognizes gestures that are required to control MS PowerPoint presentation. For this purpose, initially, the system focuses on three points, the user's head, left hand and right hand. It tracks the movements of both hands in reference to each other and in reference to the head point.

When the user moves their right hand from right to left, MS PowerPoint slideshow moves one slide forward (the current slide moves left) and the result is showing the next slide in the slideshow. If the user moves their left hand from left to right, the slideshow moves one slide backward (the current slide moves right), that is, showing the previous slide in the slideshow. This implementation was adopted as the most suitable one after several trials and experiments. An older version of the system was developed to depend only on the right hand. In that case, when the user was trying to move one slide forward by moving the hand from right to left, the system responds correctly. However, when the user returns their hand back to right, the system was wrongly recognizes the movement as a command to move the slide backwards again. The problem was handled, firstly, by taking time required for each movement into account. Reference [7] introduced that time is one of the factors that is used for controlling gesture recognition. In TeachMe, the case was that if the movement is quick, this means to perform the command while slow movement of the hand means to ignore the gesture. It was noticed that this puts some limitations on the user movement, which contradicts with the purpose of the system. Accordingly, the solution of recognizing the movements of two hands was adopted.

One of the earliest problems faced the research and the implementation of TeachMe was that the system limits the lecturer movements. This means that the lecturer once recognized by the system, they cannot move their body freely during the lecture as their gestures will be analyzed and recognized by the system even if the gestures are directed to the students not to the system. To increase the system smoothness and gives the flexibility to the lecturer to move their hands during the lectures for purposes other than controlling the slideshow or the mouse, some other factors were added to control the gesture recognition. One factor, again, is the time. From experience gained in the older version of the system, time can play a major role in gesture recognition. The time required to perform the gesture was taken into consideration. During the system testing, it has been noticed that the user moves a hand from one side to another with a reasonable speed. The user focuses on the gesture and tries to make it recognizable for the system. This makes the user uses a reasonable speed that is neither very quick nor very slow. The reasonable speed was determined subjectively by five persons who were working on the project and implemented as a factor in the gesture recognition system. Note that the system does not recognize the start of the gesture. Instead, the system knows the gesture when it ends. Based on this, the system measures the time required for any movement of the user's body (including hands) and sends the data to be recognized as gesture or not.

Another factor taken into consideration was the second hand position while one hand is making the gesture. Moving right hand from right to left is not enough to generate the command of moving the slides forward, or "show next slide", in TeachMe. Instead, left hand should be in the left part and below the head point while the right hand is moving from right to left. Note that right and left sides of the body are recognized based on the head point that is considered by TeachMe implementation, as the "middle of the body" point reference.

One more factor that was added as a feature to recognize is the distance a hand moves from one side to another. The initial and the final positions of a hand are captured. A minimum threshold of 25 cm is set. Using threshold helps the lecturer to move their hands freely while lecturing without generating undesired commands because threshold adds more features to satisfy a gesture recognition. Threshold is introduced by [2] to recognize gestures. As a summary, the hand position according to the head point, the second hand position, the distance of the movement and the time required to go from first point (position) to the final point are all combined to recognize a gesture and generate a command. In general, it has been noticed that adding more features helps in better recognition for the required command and gives more freedom and flexibility to the user to complete their lecture smoothly.

In addition to moving slides forward and backward, TeachMe enables the user to start and end a PowerPoint presentation. Starting a presentation is done through moving the right hand side from down to up. The hand should start

moving from a point below the head to end up above the head point. Moving the left hand similarly ends the presentation. In both cases, the other hand should be in its normal relaxing position referring to the head point. Similar to previous movements, speed and distance of the movement are all required to recognize the start and end presentation commands. Apparently, TeachMe depends on a set of features that should be tested and satisfied before a gesture is recognized.

Reference [7] introduced an approach to recognize gesture using gesture description language. This language depends on text script written as rules with an expert system approach to manage the rules and recognize the gesture. The above-described way of implementing rules in TeachMe is a similar approach followed by [7] as each rule defines specific positions for joints. Rules implemented in this research are simpler, however. Reference [7] took into consideration the distance between joints as a factor for gesture recognition as this research does. Although they mention that time is an important factor in gesture recognition, they used the time in segmentation of the captured images not in the rules. On the contrast, TeachMe takes time directly as a factor for gesture recognition. Instead, they depended on time sequence for the poses to form a gesture. They say "our research presented in this paper proves that it is possible to unambiguously recognize, in real time (online recognition), a set of static poses and body gestures (even those that have many common parts in trajectories) using forward chaining reasoning schema when sets of gestures are described with an ''if-like'' set of rules with the ability to detect time sequences". The if-like in their implementation is similar to the thresholds considered in this research.

IV. Controlling Mouse Pointer

TeachMe requires two hands to control the mouse pointer. The user should raise their left hand to be approximately beside the head point and keep it fixed in that position. A 5 cm threshold gap is set between the left hand position in reference to head point. The user uses right hand to control the mouse pointer movement by placing the right hand in front of his/her body and moves the hand. The mouse pointer moves with the hand movements. Mouse pointer control is possible using the same gesture on Windows desktop or any other running application.

To generate a left mouse click command, selecting an icon on the desktop as an example, the user should move the mouse over the required item to be selected, lower down the left hand side to be under the head point, which prevents the mouse pointer movement, and then, move their left hand forward away from the body. This hand push movement uses the Z-axis. There should be a specific distance between the body point and the hand point in Z-dimension to recognize the gesture as a click. Of course, there is a space threshold set here for the distance in Z-dimension. Double click command is achieved when the user repeats the forward movement twice with the left hand. The time between the two movements is one of the major features to recognize the gesture as a one click or double click. If the user leaves a significant period of time, between the two movements, the gesture will be recognized as two gestures of a click which may lead to a different behavior. Traditional mouse uses the same concept to differentiate between the one click and double click. The period between the two movements was set to be less than two second to be recognized as a double click.

TeachMe also allows the user to move objects in the desktop using the mouse pointer. The user may use his/her both, left and right, hands in a push gesture in front of the body that leads to a holding of a left mouse click behavior. In this case, the gesture uses the z-dimension but for both hands this time. When the user moves his/her both hands together, the mouse pointer moves accordingly. Using this gesture, the user can drag and drop folders into others or into the recycle bin as an example. The gesture depends on the keeping the two hands beside each other and in front of the body with a distance at the same time.

V. Gesture-Command Customization

TeachMe can be customized to suit the user preference. The customization tool is implemented to help the user specify a specific gesture to a specific command. In TeachMe current implementation, the commands that can be customized are those controlling a PowerPoint presentation but not the mouse pointer controlling. Once the system starts, the above described gestures and related commands are set as default for the user. If the user requires to modify a specific gesture to perform another command, customization tool can help. The tool shows four commands that can be customized, moving a slideshow forward, moving a slideshow backward, starting a slideshow and ending it. As an example, if the user wants to customize moving a slideshow one slide forward using the gesture "moving right hand from right to left" instead of "moving left hand from left to right", the default, they can specify this by selecting the "Moving Forward" command using a radio button. This opens a dropdown list with possible gestures from which the user can select, "Right" in this case. To continue the example, the user also customize "Moving Backward" command to be "Left". This customization appears in (Fig. 1). Similarly, the user can customize gestures for starting and ending slideshow commands (Fig. 2 and Fig. 3). According to the literature reviewed, such customization feature has not been implemented at any system before. The following figures show how the user can customize the system.

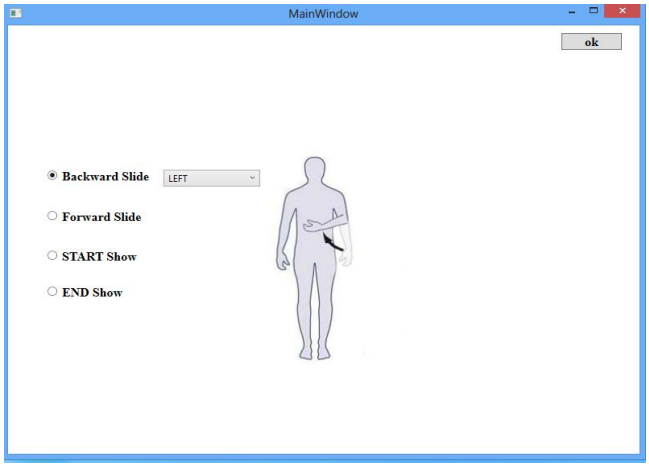


Fig. 1. Customizing moving a slide backward gesture

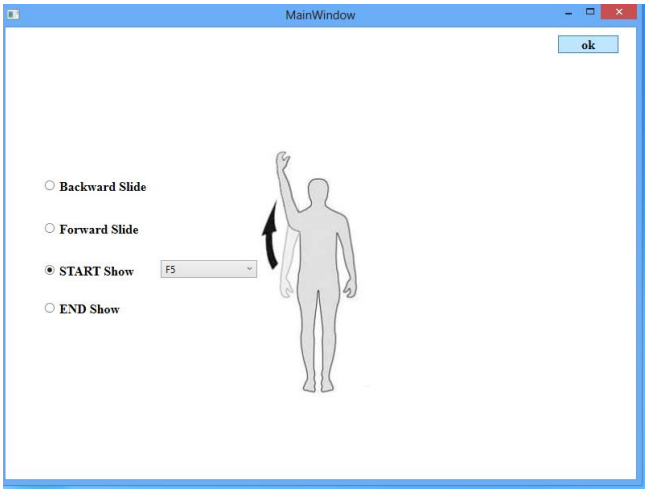


Fig. 2. Customizing starting a presentation gesture

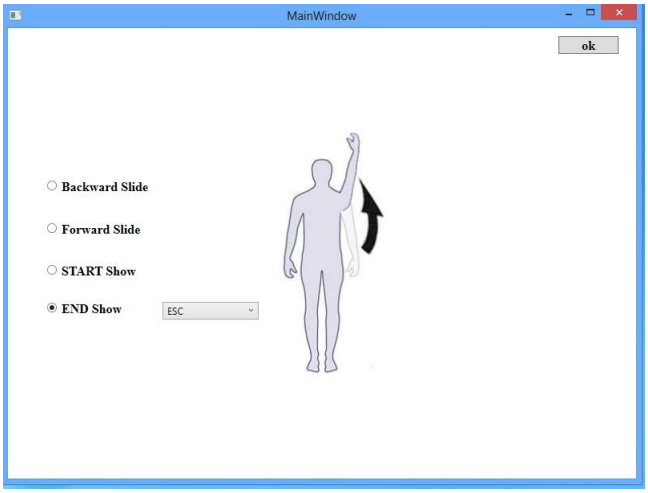


Fig. 3. Customizing ending a presentation gesture

VI. Related Work

The most related work between reviewed papers was [9]. They invented an algorithm called uWave that needs only one time training for each gesture, which provides the user the ability to personalize gestures. Their implementation and recognition depend on quantization and dynamic time warping. Implementation of TeachMe is much simpler and has a simpler customization feature application. Throughout their paper, they try to reduce the complication calculations that result from the algorithm complexity.

Reference [1] presents a gesture recognition system. Their system uses what is called Kinect Presenter approach for controlling MS PowerPoint presentation. They present an empirical evaluation for their system. The system cannot control mouse pointer nor it can control starting and ending a presentation tasks that TeachMe does. Reference [10] introduced a gesture system that controls the mouse pointer. However, it was not controlling slideshow, as it is not implemented to work in the field of education.

A vision-based gesture recognition prototype has been introduced by [11]. The system controls mouse pointer to deal with tasks related to WIMP (Windows, Icons, Menus, Pointers). The system is different from TeachMe that it does not control presentations and it is not designed for presentation purposes.

The research of [2] was motivated, similar to this research, with the trial to find a replacement technique for the traditional, keyboard and mouse presentation controlling technique. Their way of introducing the solution was through Ki-Prez, a gesture recognition system. Their system uses Kinect for capturing gestures. However, Ki-Prez only recognizes two gestures, swiping right and swiping left that move slides forward and backward. Similarly, [12] and [3] proposed the idea of replacing the traditional techniques with gesture recognition systems using Kinect in the domain of education. Reference [12] described using such systems as an introduction for the e-learning in the future. Reference [12] introduced a prototype that allows personalizing gestures. However, they did not include technical information about the personalization technique or its implementation.

Reference [13] introduced a gesture recognition system that controls presentation. Their system characterized with ability to recognize voice in addition to gesture to refine the response of the system with the required action. Their system is distinguished from other systems with this voice command feature. However, they do not include the option to control mouse pointer into their system as voice and gesture does not leave a space for the need of mouse controlling technique.

VII. TeachMe Limitations

The main limitation of TeachMe current implementation is the limited number of gestures that can be customized using the customization option. As shown above, the user can choose to customize limited number of gestures and the options of alternative gestures are also limited. Additionally, the user can only customize PowerPoint controlling gestures but not the mouse pointer controlling gestures. It has been decided to work on solving these limitations in the next TeachMe implementation. This is clarified in the future work section of this paper.

TeachMe depends on few number of joint points, three points only, for gesture recognition, reduces its ability to recognize complex gestures. However, from another point of view, using three points only to recognize gestures makes recognition process easy and reduces its complication. Putting in mind that TeachMe is a prototype gesture system designed for research, it was decided to make it simple to concentration on the new feature implementation instead of recognizing complicated gestures implementation. In addition, TeachMe satisfies all the research needs regarding the gestures recognition, as the research requires only simple gestures.

VIII. Future Work

The following work will be evaluating the new customization feature empirically. This will give an indication of continuing work in this field or find a diversion to another track. One more future work is to enhance the customization feature in TeachMe. Enhancement thought of is to introduce programming by example concept in customizing TeachMe. This will increase the flexibility of customizing gestures and

increase the number of gestures that can be customized according to the user preference.

IX. CONCLUSION

This paper introduced a gesture recognition system, TeachMe. The system recognizes gestures that control MS PowerPoint presentation and mouse pointer. It depends on Kinect as an interface for capturing gestures. The paper discussed the technical implementation details of the system. TeachMe implementation includes a gesture customization feature that allows the user to customize gesture according to their preference. Examples showing the work of the customization feature were included.

ACKNOWLEDGEMENTS

We would like to thank the students who participated in building TeachMe prototype as part of their graduation project, Mohammad Zedan Ismael and Rslan Zedan. The authors are grateful to the Applied Science Private University, Amman, Jordan, for the full financial support granted to this research.

REFERENCES

[1] S. Cuccurullo, R. Francese, S. Murad, I. Passero and M. Tucci, "A gestural approach to presentation exploiting motion capture metaphors," in *Proceedings of the International Working Conference on Advanced Visual Interfaces*, Capri Island, Italy, 2012.

[2] D. Martinovikj and N. Ackovska, "gesture recognition solution for presentation control," in *The 10th Conference for Informatics and Information Technology (CIIT 2013)*, 2013.

[3] S. Butnariu and F. Girbacia, "Development of a Natural User Interface for Intuitive Presentations in Educational Process," in *The 8th International Scientific Conference eLearning and software for Education*, Bucharest, Romania, 2012.

[4] S. Change, "Using Gesture Recognition to Control PowerPoint Using the Microsoft Kinect," MIT, Massachusetts, 2013.

[5] S. Ha, S. Park, H. Hong and N. Kim, "Study on Gesture and Voice-based Interaction in Perspective of a Presentation Support Tool," *Journal of the Ergonomics Society of Korea*, vol. 31, no. 4, pp. 593-599, 2012.

[6] T. Osunkoya and J.-C. Chern, "Gesture-Based Human-Computer-Interaction Using Kinect for Windows Mouse Control and PowerPoint Presentation," in *Proc. For the 46th. Midwest instruction and computing symposium (MICS2013)*, Wisconsin, 2013.

[7] T. Hachaj and M. Ogiela , "Rule-based approach to recognizing human body poses and gestures in real time," *Multimedia Systems*, vol. 20, no. 1, pp. 81-99, 2014.

[8] A. Butalia, D. Shah and R. Dharaskar, "Gesture recognition System," *International Journal of Computer Applications*, vol. 1, no. 5, p. 48–53, 2010.

[9] J. Liu, L. Zhonga, J. Wickramasuriya and V. Vasudevan, "uWave: Accelerometer-based personalized gesture recognition and its applications," *Pervasive and Mobile Computing*, vol. 5, no. 6, p. 657–675, 2009.

[10] A. Argyros and M. Lourakis, "Vision-based interpretation of hand gestures for remote control of a computer mouse," in *Computer Vision in Human-Computer Interaction*, Berlin Heidelberg, Springer, 2006, pp. 40-51.

[11] F. Farhadi-Niaki, R. GhasemAghaei and A. Arya, "Empirical Study of a Vision-based Depth-Sensitive Human-Computer Interaction System," in *10th Asia Pacific Conference on Computer Human Interaction*, Matsue, Japan, 2012.

[12] V. Tam and L.-S. Li, "Integrating the Kinect Camera, Gesture Recognition and Mobile Devices for Interactive Discussion," in *IEEE International Conference on Teaching, Assessment, and Learning for Engineering (TALE)*, Hong Kong, 2012.

[13] J. Kim, S. Kim, K. Hong, D. Jean and K. Jung, "Presentation Interface Based on Gesture and Voice Recognition," in *Multimedia and Ubiquitous Engineering*, Berlin Heidelberg, Springer, 2014, pp. 75-81.

A Mobile Device Software to Improve Construction Sites Communications "MoSIC"

Adel Khelifi
College of Computer Information Technology
American University in the Emirates
Dubai, UAE

Khaled Hesham Hyari
Civil Engineering Department
Hashemite University
P.O. Box 330127, Zarqa 13133, Jordan

***Abstract*—Effective communication among project participants in construction sites is a real dilemma for construction projects productivity. To improve the efficiency of participants in construction projects and have a speedy delivery of these projects, this paper presents the development of a mobile application system to support construction site communication. The developed system is designed to enhance communication between home office employees, field office staff, and mobile users at the construction sites. It has two components: a mobile application and a website. The mobile application component provides users with valuable features such as, receive sites' instructions, send requests for interpretations and retrieve information about projects. Whereas, the website component allows users, such as, home office employees to track projects' progress and find projects' location. The developed system is tested first on emulators and then on Android devices. After that the system was tested on a highway improvement project. Through their mobile phones, site users are able to interact with field office and home office personnel who use the web application to communicate with mobile users. It is expected that this work will contribute to facilitate communication in construction sites, which is much needed in this information intensive sector.**

Keywords*—Mobile application ; Construction communication; Construction site ; Construction information

I. Introduction

The nature of construction projects present unique communications challenges among project participants. These challenges can be attributed to: (1) The massive amount of information that needs to be transferred and exchanged during the construction phase of projects as it is well known that construction is an information intensive industry [1, 2, 3]; (2) The spatial dispersion of project teams and construction activities as well as frequent changes of work site locations [2, 4]; (3) The fragmented nature of the industry that engage many different stakeholders from the owner, consultants, bankers besides contractors, subcontractors and suppliers which creates gaps in information flow [3, 5, 6 ,7]; (4) The separation between site offices and work sites [1]; (5) The need for timely transfer of information as the construction industry is characterized by rigid deadlines and costly delays [8]; and (6) The increased reliance on subcontractors to perform construction work as it was reported that at least 80% of the activities performed on a typical construction site is subcontracted to specialty contractors which in turn deepen the fragmentation of the industry [6].

In 1990, John Holingworth, the head of IT at Wimpey construction said "The construction industry is as much a manager of information as it is of materials." This statement was part of his contribution to the "Building IT 2000" organized by the Building Centre Trust in 1990. This is now a lot truer with the wide use of internet and computing in all business environments [9]. Effective collaboration and communication among project participants is crucial for successful construction projects [10]. Accurate and reliable information should be exchanged in a timely manner to make informed decisions as well as monitor and control construction projects [11].

Several researchers indicated that ineffective communication practices are one of the serious factors that contribute to the poor performance of the construction industry in terms of low productivity that leads to higher project cost and time overruns [7, 8, 12, 13]. Mohamed and Stewart [8] indicate that cost overruns and delays in construction projects can be attributed to inferior coordination resulted from poor information handling and exchange, inadequate, insufficient, inaccurate, inappropriate, inconsistent, late information or a combination of them all. Meland et al. [14] indicate that empirical research has identified poor communication and information logistics as one of the central determinants of project failure.

There is a need to support communication among project' participants during the construction phase of projects. The objective of this paper is to present the development of a mobile application for construction site communication. The developed system is expected to improve the efficiency of information transfer and exchange in the construction sites. This in turn will help to improve productivity at construction sites and contribute to timely delivery of construction projects.

This paper is organized in six sections as follows: after the introduction section, the literature review section below provides a summary of previous research efforts that addressed mobile computing in construction projects. Section three discusses the system's design. The details of the system implementation and testing are described in section four. Section five presents the conclusion and future enhancements of this project.

II. Literature Review

The critical role of information management in the success of construction site administration has motivated construction

scholars to investigate the opportunities brought by the fast developments in mobile computing to promote effective communications in construction projects. Construction scholars have recognized the promises of mobile computing and how it can improve the efficiency of construction site operations [12, 15, 16]. Mobile computing came as a perfect match to the construction industry that is characterized for a long time as an industry with portable plant and mobile work force. Mobile workforce and site offices badly need a way to access information systems and exchange updated information to enable timely decisions and efficient documentation of work activities. Mobile computing quickly started to be a major theme in construction management research [1]. Research in this area can be grouped into two major categories: the first one addresses the status and prospects of utilizing mobile technology in the construction industry as well as industry needs and challenges faced [11, 15, 17]. The second one focuses on reporting the development of mobile computing tools and their implementation at construction sites [4, 18].

Atalah and Seymour [11] surveyed the current state of wireless information technology in the construction industry and reported that the level of interest in wireless technology is much higher than the level of use. About 60% of respondents to a web-based survey indicated high interest in mobile applications that can complete daily reports, safety checklists, and quality checklists in the construction sites. Chen and Kamara [1] indicate that although mobile computing in construction is a major research theme and hot research area, most of the reported research in this area focuses on a detailed aspect or single component of a mobile computing technology. Kim et al. [19] presented a location-based construction site management system using a mobile computing communication platform. The system includes a site management module and a construction drawing sharing module. The first module provides the location information of both construction activities and the resources allocated to the activities, while the second module provides an easy access to construction drawing. Venkatraman and Yoong [4] described the development of a mobile facsimile solution called "ClikiFax" that can assist collaborative communications between parties on or away from the construction site, and reported that field testing and evaluation of the tool indicated its usefulness in conveying dynamic changes to site drawings and approvals at remote construction sites. Löfgren [12] presented a case study of utilizing mobile computing by one of the largest construction companies in Sweden to manage construction site operations in a pilot project. The study concluded that wirelessly connected tablet devices on site added new functionality and flexibility to the existing fixed communication infrastructure and information systems. The users of the system experienced better productivity and faster resolution of problems due to better communication between project management and production staff in the construction site.

Jadid and Idress [16] presented an approach for utilizing mobile computing in the construction sites based on wireless LAN network and the use of personal digital assistants (PDAs) in construction site. Kimoto et al. [18] developed a PDA-based mobile computing system for construction managers at construction sites. The system includes two components: the data input program in PDA and the output program in PC. The system provides the capability of accessing checklists and reference data such as specifications and drawings. Beyh and Kagioglou [2] investigated the potential use of IP telephony (Internet Protocol telephony) to facilitate communications at construction sites. The IP telephony can be used to exchange voice, facsimile, and/or voice-messaging applications that are transported via the Internet, rather than the Public Switch Telephone Network (PSTN). The main disadvantage of the proposed IP telephony is totally dependency on local area network (i.e. Wi-Fi internet connectivity). Also, it is not possible to identify the exact geographic location of the given IP address in case of emergencies whereas this is possible using mobile phones or the public switched telephone network system.

While previous research efforts have made considerable advancements to construction site communications, it should be noted that most of the developed mobile computing solutions are based on the availability of wireless local area networks (i.e. Wi-Fi communication) throughout the construction site. This requirement is reported as a communication challenge or barrier as it is difficult to provide and maintain wireless networks considering the nature and environment of construction sites (dust, concrete walls and slabs, huge construction sites, mobile work zones). There is a need for a new application that overcomes the above limitations. The remaining sections will present the development of a mobile application based on 3G communication to support communications at construction sites. The development of the mobile application includes the design of the system and its implementation.

III. System Design

The present model utilizes the third generation of mobile telephony (i.e. 3G communication technology) while previous applications are based on wireless local area networks (i.e. Wi-Fi communication). Although both technologies provide wireless internet access services to the users, the major difference is the methods of connection to the internet. Wi-Fi connects to the internet through wireless network, and is characterized by its short range. The range of service depends on the proximity to the router. Currently, Wi-Fi is provided through private networks at homes and offices, or in public spaces like shopping centers, airport. This method has limited application in construction sites, especially mobile work zones which limits its usage. On the other hand, 3G is a type of cellular network and connects to the internet wherever there is mobile phone service. This means its range is a lot wider than a wireless network. As such the present model is developed as a mobile application based on 3G technology. The design of the system includes two major steps: (1) requirement gathering and analysis; and (2) developing the use case diagram. The following subsections provide a brief description of those steps.

A. The Software Requirements Gathering and Analysis

The performed requirement analysis is intended to identify the essential system requirements from a user's point of view. The performed analysis provides detailed information about the

user's objectives and the needed features of the system. The identified requirements include functional and nonfunctional requirements. Functional requirements describe what the system should do, while non-functional requirements describe how the system works [20]. Appropriate requirement analysis can reduce the effort wasted in upgrading, recoding and retesting the system. Although construction communication is complex, the identification of communication system requirements in construction projects is attainable due to the large number of research publications that addressed the nature of communication in the construction industry [5]. The functional requirements of the construction site communication system include both: (1) web application requirements; and (2) mobile application requirements. Table 1 illustrates the functional requirements for the developed system as well as the actors involved in performing these functions. The non-functional requirements for the proposed system consist of: (1) runtime system qualities; and (2) other system qualities. The runtime system qualities include: availability, usability, performance, security, reliability. Examples of other system qualities include: modifiability and portability.

TABLE I. FUNCTIONAL REQUIREMENTS FOR THE DEVELOPED SYSTEM

Functional Requirement	Actors Involved	Description
Register in the system	Mobile User	Allow the mobile user to register in the MoSIC and start using its features.
Login/Logout	All Users	Allow the user to access the MoSIC system by entering his/her usernames and passwords. Each type of users will have a different home page depending on his/her privilege.
Add Projects	Home office employee	Allow the home office employee to create new projects.
Modify Projects	Field office employee	Allow the field office employee to modify the content of an existing project.
Delete Projects	Home office employee	Allow the home office employee to delete the project that has been created before.
Add Reports	Mobile User	Allow the mobile user to add reports to the existing projects.
View Projects	Home office employee	Allow the home office employee to view a specific project or all projects.
View Projects Location	Home office employee	Allow the home office employee to view the physical location of the project.
View Activity Location	Mobile user, Field office employee	Allow the mobile user/field office employee to view the physical location of any activity in the project.
Capture Pictures	Mobile User	Allow the mobile user to take a picture of project activities and send it to other users.
Send PDF Reports	Mobile User	Allow the mobile user to send reports of selected project activities.
Call for meetings	Field office employee	Allow the field office employee to call mobile user/s for meetings.
Tracking Project Progress	Home office Employee	Allow the home office employee to view the progress of the existing projects.
Tracking Activity Progress	Field office Employee	Allow the field office employee to view the progress of the existing projects.
Assign Privileges to users	System Administrator	Allow the system administrator to assign privileges to all users

B. The Software Use Case Diagram

The use case diagram is a graphical representation of the interactions among the elements of a system. It describes how a user uses a system to accomplish a particular goal. A use case acts as a software modeling technique that defines the features to be implemented and the resolution of any errors that may be encountered. The components of a use case diagram include: (1) the boundary, which defines the system of interest in relation to the world around it; (2) the actors, usually individuals involved with the system according to their roles; (3) the use cases, which are the specific roles played by the actors within and around the system; and (4) the relationships between and among the actors and the use cases. As shown in Figure 1, the present system is designed to support four types of users:

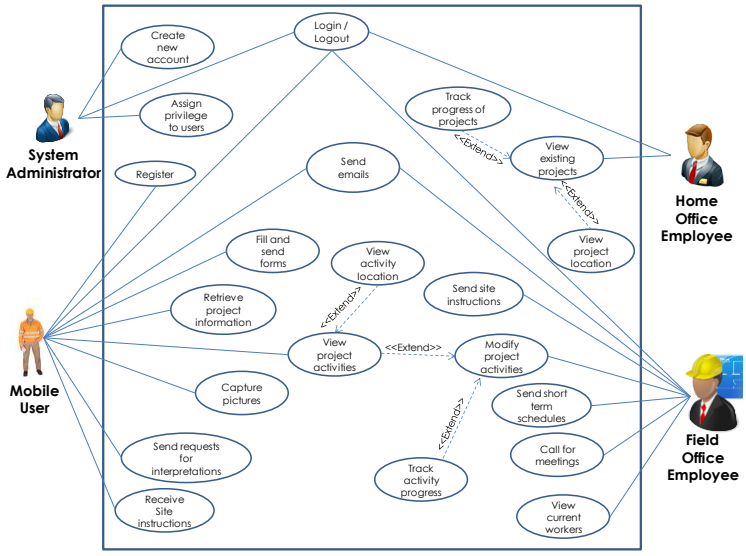


Fig. 1. The Software Use Case Diagram

(1) Home office employees which include the management staff at the home office. Through the web application, home office employees will be able to communicate with construction site managers at the field offices through the web application, and they will be able to communicate with construction site supervisors in all projects that are using the MoSIC. Home office employee can: (i) view all existing projects; (ii) track progress of any project within the system; (iii) view physical location of any project within the system; (iv) view all communications between construction

site managers and their site teams (i.e. mobile users); and (v) send emails to construction site managers and receive emails/pictures /reports from them.

(2) Field-office employees which include construction site manager, superintendent, and site office personnel. Through the web application, construction site managers will be able to communicate with MoSIC users at the construction site (i.e. supervisory staff and crew managers). When the site manager is logged into the system, he/she can: (i) view all existing activities in the site; (ii) send site instructions; (iii) send short term schedules to construction crews; (iv) send email to a specific supervisory staff at the site; and receive emails/pictures/reports from them; and (v) call for meetings. When the field office employee chooses a specific activity after selecting "View Project Activities", he/she can (a) add/delete/modify activities; (b) track activity progress; (c) view activity physical location; and (d) view all current workers on any selected activity.

(3) Construction site users or mobile users which include supervisors, crew managers, and quantity surveyors: After downloading the mobile application (MoSIC), each site user needs to register in the system in order to create an account in the database for that user. Afterward, mobile users will be able to communicate with field office management staff and access project information. Through their smart phones, mobile site user can: (i) view project activities; (ii) send requests for interpretations or design clarifications; (iii) retrieve project information related to the task under consideration (e.g. construction drawings, specifications, quality control checklists, safety auditing checklists, construction method statements, and materials tracking information); (iv) receive site instructions through the construction site manager; (v) fill and send forms related to performed tasks (e.g. daily job logs, safety incident reports, testing and inspection reports, measured quantities of work, equipment management forms); (vi) capture pictures documenting actual circumstances and send them to field or home offices as a reference to a specific issue; and (vii)) access and update punch-listing. When the mobile user chooses a specific activity after selecting "View Project Activities", he/she can (a) view activity physical location; (b) look at activity scheduled times, and (c) see activity allocated resources.

(4) System administrator who can create new accounts for users, and assign privileges to users. System administrator will be within the home office management staff. Home office employees will have the privileges of creating new projects, deleting projects, and accessing data related to all projects. Field office employee will have the privileges of accessing his/her project and adding/deleting/modifying activities in the project. On the other hand, mobile users will have the privilege of viewing project documents, and uploading documents but they are not allowed to delete documents.

IV. System Implementation

The implementation stage is the phase where software developers create the code to have a true application that fulfills all prerequisites pointed out in the past stage. The principal phase of implementation incorporates two major steps. Making a high level design is the first step, where more specific designs are set that fulfill the pre-stated system requirements (i.e. functional and non-functional requirements). Such designs are developed for the use of system's programmers rather than its clients and can be considered as implementation plan. These designs incorporate system architecture, graphical user interface, software components and data storage designs.

Coding is the second step in the system's implementation. The programmers take the software framework, its architecture, and its detailed design; define their programming environment and tools (i.e. editors, compilers and debuggers); then they start writing the code. The following section presents the implementation components of the present system: 1) the system architecture; 2) the website implementation and testing; 3) the mobile application implementation and testing; and 4) the interface between mobile application and the website.

A. The Software Architecture

An architecture design is a plan for how the system will be distributed across the information technology devices, network environment and what hardware and software will be used for each device. As illustrated in Figure 2, the client-server architecture is selected to implement the model. Mobile application with Web based systems usually follow this architecture, with the mobile phone device (the client) performing presentation and only minimal application logic using programming languages such as Android Stack, while the server has the application logic, data access logic and storage.

Fig. 2. The System Client-Server Architecture

The benefits of client-server architecture are summarized in two points. First of all, it allows for scalability. This means that it is easy to increase or decrease the storage and processing capabilities of the servers. If one server becomes overloaded, you simply add another server so that many servers are used to perform the application logic, data access logic, or data storage. Second, client-server architecture can support many different types of clients and servers, meaning that it is possible to connect mobiles and computers that use different operating systems. In addition, for thin client server architecture, ones that contain a small portion of the application logic, that use the internet standards, it is simple to clearly separate the presentation logic, application logic and data access logic and design each to be somewhat independent. This means that the interface can be changed without affecting the application logic and vice versa. Therefore, the concept of client-server architecture was implemented to be MoSIC's architecture as shown in Figure 3.

The developed system has three layers, which are: (1) presentation layer, (2) application layer; and (3) data access layer. The first layer is the presentation layer. It presents project information, such as construction drawings and methods of construction. It displays forms that to be filled by users, such as daily job log, equipment management, measurements of actual work performed, incident report and inspection results. Also, it allows sending and receiving text report, such as sending request for interpretations and receiving site instructions. In addition, this layer displays projects' location on digital maps. The application layer is where the software and hardware requirements are integrated to meet the application specifications. In this layer, mobile computing infrastructure components such as camera, digital maps, wireless technologies, GPS, internet, touch screen and mobile device are all collaborating within the application.

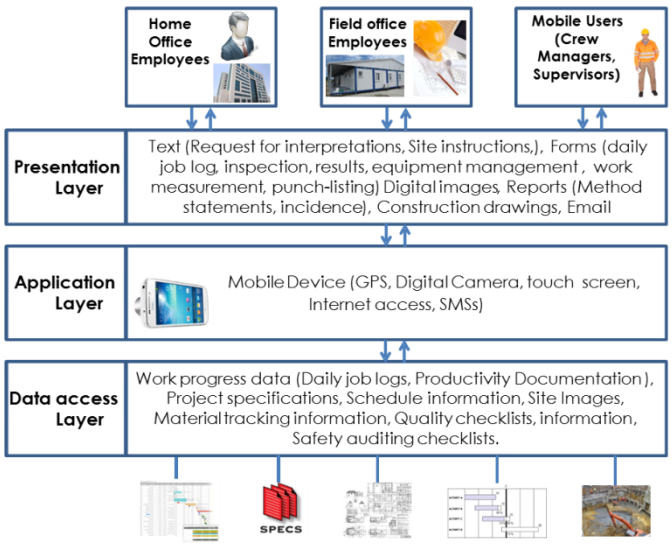


Fig. 3. The Mobile Devices Software Layers

The last layer is the data access layer. It stores projects' information, such as construction schedules, project's technical specifications, material tracking, daily job logs, safety and quality checklists, work progress data, construction site image data. Information modifications are well recorded in this tier.

New recorded information is automatically sent to the database server via a fourth or third-generation wireless standard for mobile devices. According to the above system architecture, the system was developed using Android software stack. It is Java and the new Android Studio, along with the Android Software Development Kit (SDK).

B. Website Implementation and Testing

The MoSIC website was developed over six months and runs on the IIS (Internet Information Server). C Sharp (C#) programming language is used to develop the website. C# is one of ASP languages. ASP.NET stands for Active Server Pages.NET. It is a web application framework which is developed by Microsoft in January 2002 in order to help programmers in creating data-driven websites using the .NET platform. MoSIC website is developed using ASP.NET technology since it has many advantages. Kozyk [21] highlighted few of these advantages "ASP.NET provides better performance by taking advantage of early binding, just-in-time compilation, native optimization, and caching services right out of the box. The source code and HTML are together therefore ASP.NET pages are easy to maintain and write. Also the source code is executed on the server. This provides a lot of power and flexibility to the web pages". The website has main features for the Home Office Employees such as view existing projects, view projects' location and track projects' progress. It provides key interactive features for the system administrator, as well which are creating new accounts and assign privileges to users.

The implementation phase included two months of testing phase. Testing our website is crucial during the development phase and after the website is built. During its development, its pages were previewed periodically to make sure the webpage are working properly and the site presents the features defined in the design phase and fulfill its purpose. To do so, the expected result of each page in the website is listed. After running the website, actual and expected behavior of each page are compared. In case of mismatch, a note is added to state what is problematic and what needs to be fixed. At the end, it is ensured that the website works in significant browsers such as Google Chrome, Internet Explorer , and Firefox.

C. Mobile Devices Application Implementation and Testing

There are several possible implementation techniques for mobile applications. Each one has its advantages and disadvantages. The "native" implementation for MoSIC development is selected. A "native" implementation for mobile applications means writing the application code using the programming language and programmatic interfaces provided by the mobile operating system of a specific type of device. The present mobile application was developed as a native implementation for Android mobile devices. As such, the implementation was written using Java programming language, Eclipse integrated development environment, the Android operating system (Linux) and Application Programming Interfaces (APIs) that Android supplies and supports. The native application implementation is selected because it ensures highest compatibility with the mobile device since the programming language and APIs used are specific to the hardware for which the application is developed. Also, native application can take total benefit of every library, function or service provided by the device. However, the developed mobile application will work only on Android mobile devices. A native Apple iOS application should be completely redeveloped in order to enable the users of Apple mobile devices to use the developed system. Future developments of MoSIC will be developing a MoSIC version that will be more compatible with other types of mobile devices. During the development phase a focus was set on specific set of features that meet the identified system requirements. The application will evolve horizontally through adding more functions to it as the need arises. Figure 4 illustrates the MoSIC main features.

Testing is one of the most essential steps in building mobile applications. Testing gives accurate impression of what users are going to experience when they use the application. It allows correcting errors and application's usage issues before its deployment; therefore, developers will avoid customers'

frustrations later. The MoSIC mobile application part was tested on emulators and then on real mobile devices in order to get the real look and feel of the mobile application. First, MoSIC is tested on computer desktop emulators for Android. This means running the application on the emulators and navigates through all its screens to test them. The android emulator is available in its developer's kit. However, emulators are recommended as first step for testing mobile applications, but they are slower than the real device. In addition, they can't present all application's features. For example, emulators interact with computer mouse not with a finger like on a phone touch screen. Second, the team tested MoSIC on three different Android devices.

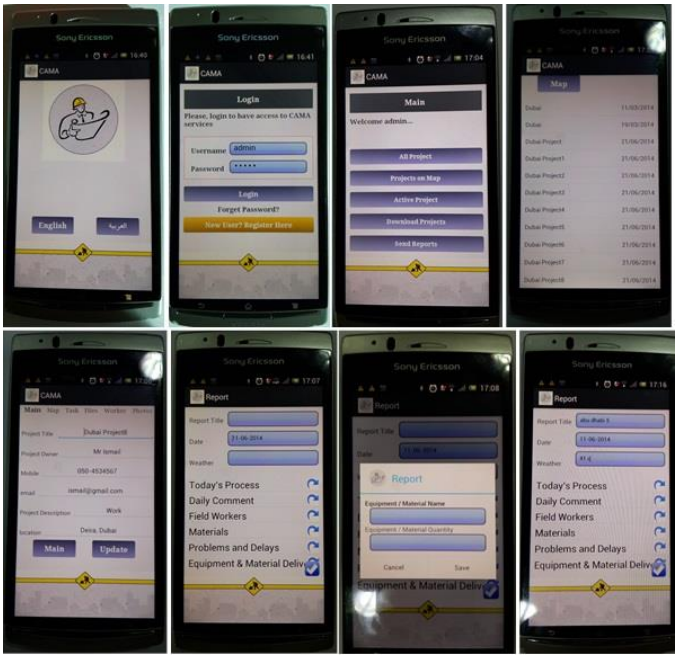


Fig. 4. Main Features of the Mobile Component of MoSIC

There are mainly two methods of software testing: White-box testing and Black-box testing. White-box testing is a software testing technique that inspects application's code, its algorithms and their efficiency as opposed to its functionalities. Black-box testing is a technique of software testing that tests functions of an application. It verifies what the application does without examining the code or its algorithm. It needs the preparation of test cases for input and expected outputs. Then expected outputs are compared to actual outputs. Any discrepancy between the types of outputs will reveal errors in the application's behavior.

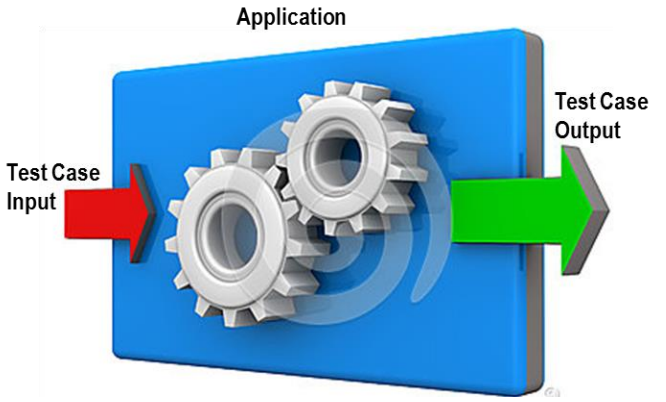


Fig. 5. MoSIC Black Box Testing

Since the native implementation technique is chosen for MoSIC, where the emphasis on developing MoSIC features, black-box testing is utilized for examining system functionalities as illustrated in Figure 5. As an example, Table 2 shows a test case's result after performing black-box testing of MoSIC mobile application login screen.

TABLE II. TEST CASE RESULT

Project Name: MoSIC						
Test Case 1 **Test Case ID:** 1 **Test Priority (Low/Medium/High):** Med **Module Name:** MoSIC login screen **Test Title:** Verify login with valid username and password **Description:** Test the MoSIC login screen				**Test Designed by:** Dr. Adel Khelifi **Test Designed date:** 14-05-2016 **Test Executed by:** Dr. Khaled Hyari **Test Execution date:** 17-08-2016		
Pre-conditions: User has valid username and password						
Dependencies:						
Step	**Test Steps**	**Test Data**	**Expected Result**	**Actual Result**	**Status (Pass/Fail)**	**Notes**
1	Navigate to login page	User= 123@gmail.com	User should be able to login	User is navigated to	Pass	
2	Provide valid username	Password: 1234		dashboard with successful		
3	Provide valid password			login		
4	Click on Login button					
Post Conditions: User is validated with database and successfully login to account. The account session details are logged in database.						

All MoSIC mobile application features were tested as per the procedures in Table 2. Furthermore, MoSIC is tested in a road improvement project in Abu Dhabi city as shown in Figure 6. Through the developed application, site personnel were able to exchange data between the job site and the main office, which has improved the efficiency of communication between different parties involved in the construction project.

Fig. 6. Project Manager using MoSIC in Abu Dhabi

D. The Interface between the Mobile App and the Website

The MoSIC mobile application doesn't connect directly to MoSIC website and its database because each component uses a different operating system. As illustrated in Figure 7, the interface between mobile application and the website is developed according to the following procedures:

1) A web service is developed to facilitate the website connection and operation with the mobile application. Our web service is developed using C# programming language and SQL statement.

2) The web service is residing on a web server and waiting for Android application to send a request to it.

3) The web service formats the request to be compatible with the database attributes and opens the database.

4) The web service provides the query received from the mobile application for the requested data.

5) The web service returns the result to the Android application in a format that it can understand and closes the database.

The testing process for the interface is quite straightforward. To check if the interface between MoSIC mobile application and its website is working properly, first the web server that contains MoSIC website and its database is running. Then, the mobile application calls functions from the mobile application that interact with the website, such as retrieve project information, receive site instructions, and send requests for interpretations. This testing step is closed after receiving all requested information from the website and its database.

V. Summary and Conclusion

The construction industry is characterized as an information intensive industry with a large number of information that should be exchanged between spatially dispersed project participants. The ever improvements in mobile phones represent a unique opportunity to improve construction communication. In this paper, the development of a mobile application system is presented, which supports collaboration between mobile users at construction sites and other project participants at the field offices as well as management at the home office. As mobile users will need to interact with the database and management at the site office and the home office, the system development includes developing the mobile and the web modules as well as the interface between the two systems.

The developed system can improve the productivity and profit of construction projects as it enables field personnel at the construction site to access and upload data when needed instantaneously without going to the field office or main office to access or report information. The system also eliminates the need to gather and convey heap of papers along the construction site. Through the website, other project participants such as field office and home office personnel are able to track the progress of construction activities and projects and corresponding with mobile users on site who are utilizing MoSIC. The MoSIC permits mobile users in construction sites to record and send various types of reports related to current construction activities, and also catch and upload snapshots. Mobile users can access data related to current activities such as method statements and specifications. All data will be put away inside the database, permitting MoSIC mobile users to get to and view what is needed in the wake of interfacing with the Internet.

The developed mobile application was developed as a native mobile application for Google's android mobile devices, and therefore it is currently usable by mobile users holding android devices. Future developments should involve developing the application for other mobile platforms such as Apple's iOS and Windows. This first version of MoSIC can be enhanced further in order to increase its acceptance within the sector of construction by adding more functions such as real-time GPS tracking. Developing MoSIC with enabled voice recognition feature that allow users to interact with the application is indeed one of the main future enhancements.

VI. Future Work

These days, there is no doubt that mobile devices are omnipresent in human being life. However, due to their resources limitations, processing complex functions rest challenging for these devices. Consequently, to deal with the concept of fault tolerance, several mobile applications depend on offloading components of their systems to distant servers on the cloud or other mobile devices. In future work, the main task will be to consider the fault tolerance of mobile devices in improving the system's design. Few researchers [22], [23], [24] went beyond the traditional cloud based solution and have suggested the virtual machine overlay concept to offload mobile devices processes to adjacent infrastructures. Basically, such solution permits processes' offload, However, it entails a complex mechanism of virtual machine and a reliable connection.

To tackle the concern of fault tolerance in the context of mobile devices running sophisticated applications, Chien-An et al. [25] presented "the first k-out-of-n framework that jointly

addresses the energy-efficiency and fault-tolerance challenges. It assigns data fragments to nodes such that other nodes retrieve data reliably with minimal energy consumption. It also allows nodes to process distributed data such that the energy consumption for processing the data is minimized". After getting the feedback from the users of MoISC current version, considering such solution in the design of a new version of MoSIC will make it more reliable.

REFERENCES

[1] Chen,Y. and Kamara, J. (2011). "A Framework for Using Mobile Computing for Information Management on Construction Sites." Automation in Construction, 20(7), 776-788.

[2] Beyh, S. and Kagioglou, M. (2004). "Construction Sites Communications Towards the Integration of IP Telephony." ITCon. 9(23), 325-344. Special Issue: Mobile Computing in Construction. http://www.itcon.org/2004/23/

[3] Kajewski, S., and Weippert, A., (2003). "Online Remote Construction Management." State-of-the-Art Report, Construction Research Alliance: Queensland University of Technology (QUT) and Building, Construction & Engineering (CSIRO), Australia. http://eprints.qut.edu.au/ 4053/1/State_of_the_Art_Report.pdf

[4] Venkatraman, S., and Yoong, P. (2009). "Role of Mobile Technology in the Construction Industry – A Case Study." International Journal of Business Information Systems, 4(2): 195–209. 10.1504/IJBIS.2009.022823

[5] Perumal, V., and Abu Bakar, A. (2011). "The Needs for Standardization of Document towards an Efficient Communication in the Construction Industry." World Applied Sciences Journal, 13(9): 1988-1995.

[6] Perdomo, J. (2004). "Framework for a Decision Support Model for Supply Chain Management in the Construction Industry." Doctoral Dissertation, Virginia Polytechnic Institute and State University, Blacksburg, Virginia.

[7] Weippert, A., Kajewski, S., and Tilley, P. (2002). "Online Remote Construction Management (ORCM)." Proceedings of the International Council for Research and Innovation in Building and Construction CIB w78 Conference, 12-14 June, Arhus, Denmark.

[8] Mohamed, S., and Stewart, R. (2003). "An Empirical Investigation of Users' Perceptions of Web-based Communication on a Construction Project." Automation in Construction, 12(1): 43-53.

[9] Sun, M., and Howard, R. (2004). "Understanding I.T. in Construction." Taylor and Francis Group, London.

[10] Yang, J., Ahuja, V., and Shankar, R. (2007). "Managing Building Construction Projects through Enhanced Communication – an ICT Based Strategy for Small and Medium Enterprises." CIB World Building Congress, Cape Town, South Africa, 21-25 May 2007 , 2344-2357.

[11] Atalah, A. and Seymour, A. (2013). "The Current State of Wireless Information Technology in the Construction Industry in Ohio." The Journal of Technology Studies, 39(1), 14-27. http://scholar.lib.vt.edu/ejournals/JOTS/v39/v39n1/atalah.html

[12] Chen, Y. and Kamara, J. (2008). The Mechanisms of Information Communication on Construction Sites." FORUM Ejournal 8, June 2008: Newcastle University, 1-32.

[13] Löfgren, A. (2007). "Mobility in-Site: Implementing Mobile Computing in a Construction Enterprise." Communications of the Association for Information Systems, 20(1/37), 1-12. http://aisel.aisnet.org/cais/vol20/iss1/37

[14] Dainty, A., Moore, D., and Murray, M. (2006). "Communication in Construction: Theory and Practice." Taylor & Francis, London and New York.

[15] Meland, O., Robertsen, K., and Hannas, G. (2011). "Selection Criteria and Tender Evaluation: the Equivalent Tender Price Model (ETPM)." Proceedings of the Management and Innovation for a Sustainable Built Environment, 20 – 23 June 2011, Amsterdam, The Netherlands.

[16] Izkara, J., Perez, J., Basogain, X., and Borro, D. (2007). "Mobile Augmented Reality, an Advanced Tool for the Construction Sector." Proceedings of the 24th CIB W78 Conference, Maribor, Slovakia, 453-460.

[17] Jadid, M., and Idress, M. (2005). "Using Mobile Computing and Information Technology in Civil Engineering Construction Projects." The Journal of Engineering Research, 2(1), 25-31.

[18] Kimoto, K., Endo, K., Iwashita, S., and Fujiwara, M. (2005). "The Application of PDA as Mobile Computing System on Construction Management." Automation in Construction, 14(4), 500-511.

[19] Kim, C., Lim, H., and Kim, H. (2011). "Mobile Computing Platform for Construction Site Management." Proceedings of the 2011 International Symposium on Automation and Robotics in Construction (ISARC), Seoul, Korea.

[20] Eriksson, U. (2012). "Functional vs Non Functional Requirements." http://reqtest.com/requirements-blog/functional-vs-non-functional-requirements/ (Accessed January 15, 2016)

[21] Kozyk, S (2011). "What is ASP.NET? -Top 12 Advantages of ASP.NET." ITegrity Group. Retrieved from http://www.itegritygroup.com/asp-advantages.aspx. (Accessed June 11, 2014).

[22] Satyanarayanan, M., Bahl, P., Caceres, R. and Davies, N., (2009). The case for vm-based cloudlets in mobile computing. IEEE pervasive Computing, 8(4), pp.14-23.

[23] Chun, B.G., Ihm, S., Maniatis, P., Naik, M. and Patti, A., (2011), April. Clonecloud: elastic execution between mobile device and cloud. In Proceedings of the sixth conference on Computer systems (pp. 301-314). ACM.

[24] Kosta, S., Aucinas, A., Hui, P., Mortier, R. and Zhang, X., (2012), March. Thinkair: Dynamic resource allocation and parallel execution in the cloud for mobile code offloading. In INFOCOM, 2012 Proceedings IEEE (pp. 945-953). IEEE.

[25] Chen, C.A., Won, M., Stoleru, R. and Xie, G.G., (2015). Energy-efficient fault-tolerant data storage and processing in mobile cloud. IEEE Transactions on cloud computing, 3(1), pp.28-41.

Framework of Resource Management using Server Consolidation to Minimize Live Migration and Load Balancing

Alexander Ngenzi[*]
[*]Research Scholar
Computer Science Engineering, Jain University, Bangalore, India

Selvarani R[**]
[**]Professor and Head
Dept. Computer Science Engineering, Alliance University, Bangalore, India

Suchithra R[***]
[***]Professor and Head: Dept.Master of Science in Information Technology,
Jain University, Bangalore, India

***Abstract*—Live Migration is one of the essential operations that require more attention to addressing its high variability problems with virtual machines. We review the existing techniques of resource management to find that there are less modeling to solve this problem. The present paper introduces a novel framework that mainly targets to achieve a computational effective resource management technique. The technique uses the stochastic approach in modelling to design a new traffic management scheme that considers multiple traffic possibilities over VMs along with its switching states. Supported by an analytical modelling approach, the proposed technique offers an efficient placement of virtual machine to the physical server, performs the computation of blocks, and explores reduced resource usage. The study outcome was found to possess potential reduction in live migration, more extent of VM mapping with physical servers, and increased level of capacity.**

Keywords—Resource Management; Live Migration; Virtual Machine; Load Balancing; Cloud Computing

I. INTRODUCTION

The introduction of the cloud computing offers a change in the process of accessing as well as retrieving the data from multiple sources of clusters spread over a large geographic area [1]. The process of virtualization has played a significant contributory role for offering both data and service availability [2]. All Virtual Machines (VMs) operate in a highly integrated process resulting in an improvement in resource utilization to cater up to the massive demands on online users [3]. Owing to expensive-characteristics of cloud-based resources, it is quite imperative to perform optimization of the resource using server consolidation. In this regard, the placement of the VM is quite important to be considered when it comes to server consolidation [4] as the inappropriate placement of VM will result in maximum drainage of resources. There are various studies e.g. [5][6] that has discussed the variability problems of the traffic are existing over VM. The prime reason for this variability is the increasing adoption of enterprises with many programs that demand consistent and reliable performance. The spiky traffic over VM will represent maximized variability that assists in implicating statistical processes to evaluate the utilization trends [7][8]. A closer look into the existing techniques shows that scheduling of usual traffic utilizes the elasticity properties of cloud, but it is necessary to meet positive dynamic demands of resources to avoid overheads [9]. Therefore, live migration policies [10] and local resizing [11] are the frequently used techniques for catering up the dynamic demands of peak traffic condition. The configuration of the VM is adaptively changed in local resizing process whereas live migration results in placing some VMs to those physical servers that are found to be idle for a certain period. Although, live migration of VM is one of the most important processes associated with VM to provide seamless service delivery, it is carried out at the cost of higher resource utilization that finally results in potential downtime of some important services offered by the associated VM. Therefore, there is a need of carrying out an investigation to explore the best possibility of resource management by evolving up with a robust solution to living migration problems in the cloud along with load balancing. Therefore, the present paper has introduced one such technique which applies an analytical modeling to maintain a better level of equilibrium between live migration and efficient resource management as well as with better load balancing to the incoming traffic. The paper is arranged as per: Section 1.1 discusses the background of the study, Section 1.2 discusses the problem identified in the study, and Section 1.3 presents a brief discussion of proposed system. Section II discusses the algorithm implementation followed by analysis of result accomplished from the study in Section III. Finally, Section IV makes some concluding remarks.

A. Background

Study towards efficient resource management over cloud environment is not a new, and there has been the various amount of work has been already carried out till date. However, we will update only the most recently explored literature published in last 5 years about resource management, live migration, and server consolidation in this section. Zhu et al. [12] have incorporated a software-engineering based technique to perform scheduling of resources. Kumar and Saxena [13] have presented a study on quantitative analysis about the migration of VM along with its associated factors. Saraswathi et al. [14] have developed a technique of resource allocation to perform selection and execution of high priority task. The review paper was agreed using time and numbers of processing elements and host number. Wood et al. [15] have presented a model of live migration using dynamic pooling mechanism. The study has also presented an optimization

principle to reduce the storage cost as well as the memory of VM. Panda et al. [16] have discussed an algorithm that targets multiple environments of cloud based on the smoothening concept. The evaluation of the study was carried out using a bigger dataset of heterogeneous types. Study towards live migration problem has been carried out by Song et al. [17] where the authors have emphasized on forwarding the memory pages to retain cost effectivity in channel capacity as well as to reduce the total time of migration. Selvarani and Sadhasivam [18] have presented a task scheduling scheme over the cloud to perform mapping of the required resources. The cost of resources, as well as performance of computation, is estimated by the presented technique, and its outcome was analyzed with respect to time and cost. Nahir and Order [19] have introduced a formal framework for load-balancing using unique management policies of VM. The study outcome was testified using overhead on the mean queue. Kao et al. [20] have introduced an involuntary decision-making technique for facilitating the better process of live migration. Taking the case study of private cloud, the authors have implemented it as an experimental prototype. The study outcome was explored with better scalability; energy saving features as well as load balancing characteristics. Wei et al. [21] have addressed the problem of resisting utilization of skewed resources over physical server using resource-based prediction approach. The study has also presented a completely new technique of resource allocation of heterogeneous types for catering up multiple demands on the cloud-based networks. Yue and Chen [22] have presented a non-probabilistic technique to address the problems of VM placement over the data centers. The study outcome has shown energy efficiency as well as leads to minimization of physical servers to approximately 20%. Caton et al. [23] have used an open-source framework that uses the potential networking attributes of the social network to carry out resource allocation in the cloud. The presented study also uses stochastic modelling of node participation process. Assessment of the server consolidation was carried out by Chang et al. [24] where the problem of selection of a precise hypervisor is discussed for specific virtualization of the server. Study on dynamic allocation of resources was also carried out by Yang et al. [25] to perform autonomous migration of the jobs among the VMs depending on the amount of load. The result was assessed using time with increasing size of problem and CPU utilization using OpenNebula. Perumal and Murugaiyan [26] have adopted an optimization technique to address the problems of VM placement and consolidation of the server. Eramo et al. [27] have presented a unique architecture to solve the problem of dimensioning of server resources using optimization technique. The study outcome was found to possess better energy saving features. Study towards live migration problem was discussed by Sarker and Tang [28] has proposed an effective scheduling of VM migration policies. Ye et al. [29] have presented a framework using profiler for the purpose of minimizing physical servers along with retention of better performance of different traffic.

B. The Problem

From the previous section, it can be seen that there are various techniques towards problems related to resource allocation, live migration, server consolidation. The technical pitfalls in majority of the approaches are as follows:

- The majority of the study is focused on increasing live migration as means of server consolidation. Unfortunately, an increase of live migration also results in performance degradation while working under constraints, which is still not addressed in the existing system.
- Only a few studies have used the potential feature of stochastic and probability theory in modeling, which could be used for better visualization of different dynamicity of the traffic. In short, traffic modeling is found not to be emphasized much.
- There is lesser extent of modeling relationship between usage of physical servers, live migration, and capacity
- The proposed study identifies the above-mentioned problems and considers to solve it by its presenting analytical modeling approach. The next section briefs about the proposed solution adopted to counter-measure the identified problems.

C. The Proposed Solution

The purpose of the proposed system is to introduce a novel framework that can perform an effective server consolidation with retention of minimized live migration of VM and increased load balancing system over data centers in the cloud environment. The present work is a continuation of our prior work being carried out [25]. The complete implementation of the proposed system follows analytical modelling approach. Fig.1 highlights the proposed scheme of [FRMS].

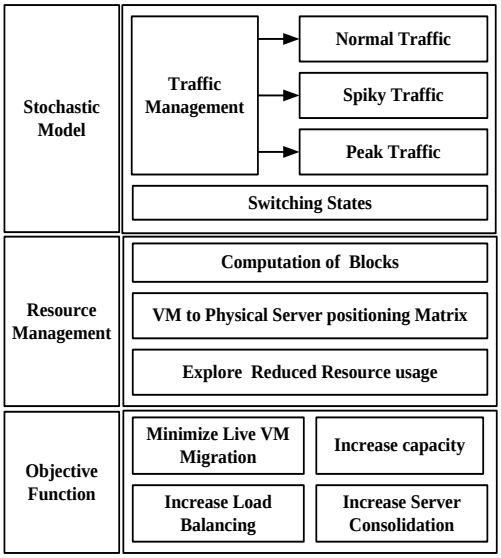


Fig. 1. Schematic Diagram of Proposed System

The proposed system introduces an empirical modeling using stochastic approach applicable for *traffic management* and *switching state* designing. The study formulates three different states of traffic i.e. *normal traffic*, *spiky traffic*, and *peak traffic* situation modeled using stochastic approach. The study also uses a probability parameter to represent its switching states i.e. states of ON and OFF corresponding to higher and normal traffic situation respectively. The resource management block mainly consists of i) *computation of blocks*

ii) *VM to physical server positioning matrix*, and iii) *exploring reduced resource usage*. The block will represent an effective serving window to perform load balancing by minimizing its number. Computation of blocks is carried out considering both the switching states with cut-off capacity value. The computed blocks assist in finding the less number of reserved spaces for physical servers. The positioning matrix assists in the allocation of VM to the respective physical serves by the number of VMs, the specification of physical servers, switching states, and capacity factor. Finally, exploration of minimal resource usage is carried out by developing a new matrix that can record only the minimal blocks needed to be allocated by the physical server on a defined spike of workload. The entire evaluation of the server consolidation is carried out by comparing mainly normal traffic and spiky traffic. Finally, the objective function is developed that is responsible for exploring mapping of VM to the physical server to minimize an event of live migration over dynamic and unpredictable traffic over cloud environment. The prime goal of the proposed approach is to ensure the existence of space with approximately zero waiting time over the load balancing system (i.e. queue). In this process, each VM that possesses its individual blocks will also be subjected to be reduced as minimal number as possible while maintaining the constraints of performance satisfied. Hence, the objective function balances minimization of live VM migration, increases capacity, maximizes load balancing system, and finally enhances server consolidation. The next section discusses algorithm implementation.

II. Algorithm Implementation

The prime purpose of the proposed algorithm is to ensure an effective resource management to be taking place in the cloud data centers with a core goal of accomplishing server consolidation. The proposed algorithm takes the input of T_n (Normal Traffic), T_v (Variable Traffic), T_h (High Traffic), η (Samples), α (number of physical servers), S_1 (Switching state-1(off→on)), S_2 (Switching state-2(on→off)), *mr* (Minimum resources), *d* (Highest number of VM permissible for physical servers), φ (capacity of host machine), τ (Capacity overflow), ρ (number of partition), N_{mig} (Number of Migration). The algorithm after processing results in live migration (denoted by N_{mig}++). The steps of the algorithms are as follows:

Algorithm for incentive allocation

Input: T_n, T_v, T_h, η, α, S_1, S_2, *mr*, *d*, φ, τ, ρ, N_{mig}

Output: Live Migration (N_{mig}++)

Start

1. init T_n, T_v, T_h,
2. $w_{load} = T_{n1} + (T_{n2} - T_{n1})$*arb(1, η);
3. for i=1:k
4. for j=1: α
5. for r=0:j
6. $op = S_1^{r(1-S_1)^{(j-r)}} . S_2^{(j-i+r).(1-S_1)^{(k-j-r)}}$
7. end
8. end
9. end
10. mr←1-[(v_1+v_2)/2]
11. for k=1:d
12. min_res←minimum_resource_block (k, S_1, S_2, mr)
13. end
14. sort(min_res)
15. for i = 1:length(PS)-1; //PS→sort(min_res)
16. X(i) = max([T_v(i),max(T_v)])*min_res(i+1)+T_n(i)+∑(T_n<φ);
17. end
18. $w_{load=}$ w_{load+1}
19. for g=2:G
20. if k ≠g
21. ρ(k)= ρ(k-1)+ ρ(k+g)
22. end
23. r = min_res(xi)*max(T_h(min(j,G)))
24. if r<r_{min}
25. r_{min}←r
26. else, N_{mig}++
27. End

End

The algorithm starts by empirically generating the traffic (Line-2). The complete algorithm performs three types of conditional checks i.e. i) of T_n=T_v, ii) T_n>T_v, and iii) T_n<T_v. Using state-based transition probability, the algorithm determines a probability factor *op* for assessing an effective load balancing (Line-6). To overcome server consolidation problem, the algorithm computes a minimum number of block

mr (Line-10), which are obtained from v1 and v2 that corresponds to the sum of all stationary distribution from 1 to *(k-1)* and 1 to *k* respectively. The stationary distribution is obtained by applying row reduction method [30]. Minimum resource block is then computed considering the input parameters of i) *k* (all numbers of VMs allowed on physical servers), ii) switching state S_1 from OFF state to ON state, iii) switching state S_2 from ON state to OFF state, and minimum resource *mr* (Line-12) that is finally sorted to obtain the better resource block (Line-14). The constraint of real Virtual Machine (VM) migration with an efficient load balancing is addressed empirically by computing X for all the physical servers (Line-16). It will mean allocation of a specific VM on the initial physical server in case the entire value obtained by X is found less than the capacity of host machine φ (Line-16) followed by incrementing traffic (Line-18). Hence, the summation of X leads to estimating a total number of used physical servers. Finally, live migration is optimized as follows viz. G is computed that represents the size of T_v is estimated (Line-19), if the number of VM permissible for physical servers (k) is not same as variable g than the algorithm empirically generates all g partitions (Line-21). Finally, minimum resource (r_{min}) is computed (Line-23) and conditionally checked to perform live migration of the VMs (Line24). Hence, as an output, it computes some live migration required to perform server consolidation.

III. Result Analysis

This section discusses the results obtained from the proposed study. The study outcome is evaluated in three different conditions of normal traffic, existing traffic, and proposed the system. The assessment was carried out for observing some used physical machines (or servers), capacity overflow, some migration, and processing time. The existing system of traffic management performs reservation of the certain specific proportion of resources on each physical server that can be considered to be permissible server consolidation strategy without any apriori of traffic.

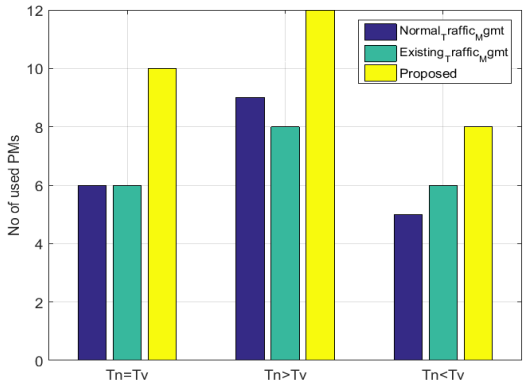


Fig. 2. Analysis of Number of used Physical Machines

The outcome in Fig.2 shows that normal traffic management uses less number of physical servers. Existing traffic management scheme is found to have similar usage of physical servers when normal traffic is equal to spiky traffic. However, in the case of difference, existing traffic management shows both lesser physical server usage (during $T_n>T_v$) and more physical server usage (during $T_n<T_v$). However, they exhibit more migration as compared to proposed system (Fig.3). This performance trend shows its attenuation pattern during the condition of $T_n>T_v$ and enhancement during the condition of $T_n<T_v$. The interesting finding is that by reducing the extent of live migration, the proposed system decreases its probability of downtime. Hence, the proposed system offers quite a less downtime and thereby exhibiting an efficient load balancing and server consolidation technique. The system, therefore, exhibits more enhanced performance by lowering down events of live migrations of VM. This outcome of lowered live migration will also have a positive impact on the capacity overflow parameter too.

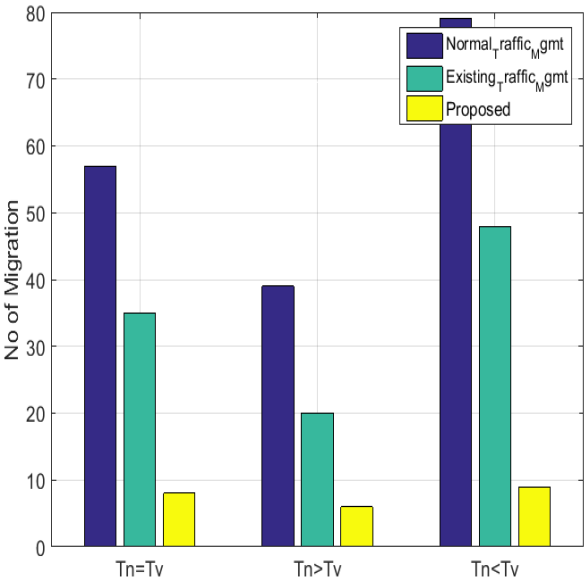


Fig. 3. Analysis of Number of Migrations

The primary intention of the proposed algorithm is to reduce the extent of the resources utilization that are kept conserved for the physical server while performing consolidation of the server and the cumulative system performance is ensured using probability theory. This will mean that a segment of time within which the collective traffic of the physical server is found to be more than its respective capacity is not higher as compared to the cut-off value of it. The proposed system applies usage of such cut-off values of capacity to resist overflow, and this phenomenon can significantly control an event of live migration thereby maintaining its capacity within very lower limits. In a nutshell, it will mean that if the capacity overflow can be controlled than live migrations of the VM can also be controlled too and hence capacity management over data center can directly influence the service quality. Fig. 4 showcases the analysis of the capacity for all the 3 different strategies.

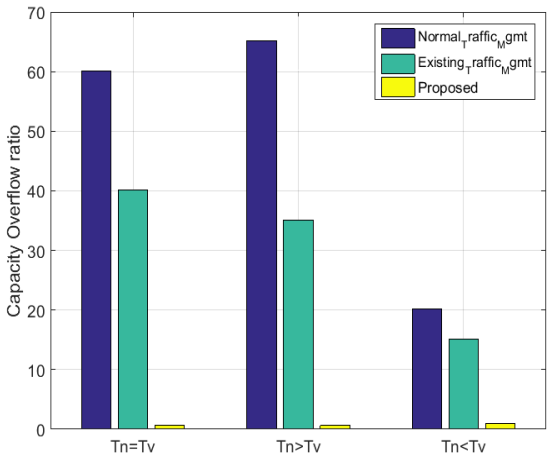


Fig. 4. Analysis of Capacity Overflow

A closer look into the graphical outcome of Fig.4 will show that proposed system to exhibit approximately 48% of enhanced performance as compared to normal traffic management and 30% improvement as compared to existing traffic management. To understand the link of capacity outcome with migrations, we consider an example where some of the physical servers may falsely declare itself as idle. It should be known that such false declaration of idle state is very common where an occupied physical server can be chosen as a target of migration. Such phenomenon will yield to the higher provisioning of physical serves resulting in iterative migration thereby causing downtime over the physical server in later stages. It also cost various resources associated with the VM to do the task scheduling under such forged cases of idleness. However, the extent of such cases is very low in the proposed system as it uses state-based transition along with probability theory that performs a minor computation of each and every resource and VMs along with its respective states. The complete testing was performed considering both stationary and changing values of traffic. To obtain convergence, the simulation study was carried out for multiple numbers of iterations individually in all the three cases of traffic pattern viz. i) $T_n=T_v$, ii) $T_n<T_v$, and iii) $T_n>T_v$.

The primary observation of the proposed system is the minimization of the number of physical servers as compared to normal traffic scene also shows more inclinations towards the adoption of such algorithms. The applicability, as well as need of such algorithm, is quite high for the massive transmission of the real-time data over the cloud. Effective allocation of the resource will further ensure a better balance between the user's request and service delivery. The outcome also showcases an effective VM migration management system along with a novel load balancing policy as well as server consolidation. Hence, a productive balance between the performance quality and utilization factor can be ensured by the proposed system. The proposed system can also be said to adopt the policy of multi-objective optimization policy where the objective function is to minimize the capacity overflow and live migration to retain a solid server consolidation scheme. The processing time of proposed technique for all the three different sates of traffic is found to be approximately 1.0576 seconds tested on core i3 machine with 64-bit Windows. When the operation environment changes than the accomplished outcome of the study only show 5% deviation as compared to stated numerical outcomes.

IV. Conclusion

With the increasing usage as well as the adoption of cloud computing, the technology consistently encounters critical challenges. One of the critical challenges that are discussed in this paper is resource management where the pivotal point of entire discussion was basically the role played by VM. In last 6 years, there has been enough number of research papers that has discussed various problems associated with VM including resource management, migration of VM, energy saving, security problems, etc. However, there is still a better scope of carrying out research work towards VM as still certain open research issues exist. The significant research issue is the lesser extent of computational modelling that focuses on the problem of live migration. There is a lot of difference between VM migration discussed in existing research work and live migration. To carry out live migration, the users will be required to be provided with the seamless delivery of services which is sustained by a higher allocation of various resources. The problems become worst if the time duration involved is more. Hence, live migration results in extensive resource usage and should be addressed properly. The proposed study presents a solution to this problem where an analytical modelling is introduced that maintains a good balance between resource management by lowering down live migration with increased capacity of VM. The outcome accomplished from the study was compared with normal traffic and existing system to find a proposed system outcomes existing system on increased use of physical servers, lower live migration, and increased capacity.

References

[1] X. Yang, "Principles, Methodologies, and Service-Oriented Approaches for Cloud Computing", *IGI Global*, 2013

[2] L. Tsai, W. Liao, "Virtualized Cloud Data Center Networks: Issues in Resource Management", *Springer*, 2016

[3] H. Saboowala, M. Abid, S. Modali, "Designing Networks and Services for the Cloud: Delivering business-grade cloud applications and services", *Cisco Press*, 2013

[4] Z. Mahmood, "Cloud Computing: Challenges, Limitations and R&D Solutions", *Springer*, 2016

[5] S. U. Khan, A. Y. Zomaya, "Handbook on Data Centers", *Springer*, 2015

[6] D. Mishchenko, "VMware ESXi: Planning, Implementation, and Security", *Cengage Learning*, 2010

[7] J. U. Gonzalez, S. P. T. Krishnan, "Building Your Next Big Thing with Google Cloud Platform: A Guide for Developers and Enterprise Architects", *Apress*, 2015

[8] C. McCain, "Mastering VMware Infrastructure 3", *John Wiley & Sons*, 2010

[9] N. L. S. da Fonseca, R. Boutaba, "Cloud Services, Networking, and Management", *John Wiley & Sons*, 2015

[10] D. Agrawal, S. Das, A.El Abbadi, "Data Management in the Cloud: Challenges and Opportunities", *Morgan & Claypool Publishers*, 2012

[11] S. Fiore, G. Aloisio, "Grid and Cloud Database Management", *Springer Science & Business Media*, 2011

[12] X. Zhu, Y. Zha, L. Liu, and P. Jiao, "General Framework for Task Scheduling and Resource Provisioning in Cloud Computing Systems", *40th IEEE Computer Society International Conference on Computers, Software & Applications*, 2016

[13] N. Kumara, S. Saxena, "Migration Performance of Cloud Applications-A Quantitative Analysis", *Elsevier-ScienceDirect- Procedia Computer Science*, Vol.45, pp.823 – 831, 2015

[14] A.T. Saraswathi, Y.RA. Kalaashri, S.Padmavathi, "Dynamic Resource Allocation Scheme in Cloud Computing", *Elsevier-ScienceDirect-Procedia Computer Science*, Vol. 47, pp.30–36, 2015

[15] T. Wood, K. K. Ramakrishnan, P. Shenoy," CloudNet: Dynamic Pooling of Cloud Resources by Live WAN Migration of Virtual Machines", *IEEE/ACM Transactions On Networking*, 2014

[16] S. K. Panda, S. Nag and P. K. Jana, "A Smoothing Based Task Scheduling Algorithm for Heterogeneous Multi-Cloud Environment", *IEEE- International Conference on Parallel, Distributed and Grid Computing*, 2014

[17] J. Song, W. Liu, F. Yin, and C. Gao, "TSMC: A Novel Approach for Live Virtual Machine Migration", *Hindawi Publishing Corporation, Journal of Applied Mathematics*, 2014

[18] S.Selvarani, G.S. Sadhasivam, "Improved Cost-Based Algorithm For Task Scheduling In Cloud Computing", *IEEE International Conference on Computational Intelligence and Computing Research*, 2010

[19] A. Nahir, A. Orda, D. Raz, "Resource Allocation and Management in Cloud Computing", *IEEE International Symposium on Integrated Network Management*, 2015

[20] M-T Kao, Y-H Cheng, and S-J Kao, "An Automatic Decision-Making Mechanism for Virtual Machine Live Migration in Private Clouds", *Hindawi Publishing Corporation, Mathematical Problems in Engineering*, 2014

[21] L. Wei, C. H. Foh, B. He, J. Cai, "Towards Efficient Resource Allocation for Heterogeneous Workloads in IaaS Clouds", *IEEE Transactions on Cloud Computing*, 2015

[22] W. Yue and Q. Chen, "Dynamic Placement of Virtual Machines with Both Deterministic and Stochastic Demands for Green Cloud Computing", *Hindawi Publishing Corporation, Mathematical Problems in Engineering*, 2014

[23] S. Caton, C. Haas, K. Chard, K. Bubendorfer, O. Rana, "A Social Compute Cloud: Allocating and Sharing Infrastructure Resources via Social Networks", *IEEE Transactions On Services Computing*, 2014

[24] B. Rong C., H-F Tsai, and C-M Chen, "Empirical Analysis of Server Consolidation and Desktop Virtualization in Cloud Computing", *Hindawi Publishing Corporation, Mathematical Problems in Engineering*, 2014

[25] C-T Yang, H-Y Cheng, and K-L Huang, "A Dynamic Resource Allocation Model for Virtual Machine Management on Cloud", *Springer Journal*, pp.581-590, 2011

[26] B. Perumal and A. Murugaiyan, "A Firefly Colony and Its Fuzzy Approach for Server Consolidation and Virtual Machine Placement in Cloud Datacenters", *Hindawi Publishing Corporation, Advances in Fuzzy Systems*, 2016

[27] V. Eramo, A. Tosti, and E.Miucci, "Server Resource Dimensioning and Routing of Service Function Chain in NFV Network Architectures", *Hindawi Publishing Corporation, Journal of Electrical and Computer Engineering*, 2016

[28] T. K. Sarker and M. Tang, "Performance-driven Live Migration of Multiple Virtual Machines in Datacenters", *IEEE International Conference on Granular Computing*, 2013

[29] K. Ye, Z. Wu, C. Wang, B. B. Zhou," Profiling-based Workload Consolidation and Migration in Virtualized Data Centres", *IEEE Transactions On Parallel And Distributed Systems*, 2013

[30] S. Andrilli, D. Hecker, "Elementary Linear Algebra", *Academic Press*, 2016

Automatic Rotation Recovery Algorithm for Accurate Digital Image and Video Watermarks Extraction

Nasr addin Ahmed Salem Al-maweri*, Aznul Qalid Md Sabri, Ali Mohammed Mansoor
Faculty of Computer Science and Information Technology,
University of Malaya, Malaysia

***Abstract*—Research in digital watermarking has evolved rapidly in the current decade. This evolution brought various different methods and algorithms for watermarking digital images and videos. Introduced methods in the field varies from weak to robust according to how tolerant the method is implemented to keep the existence of the watermark in the presence of attacks. Rotation attacks applied to the watermarked media is one of the serious attacks which many, if not most, algorithms cannot survive. In this paper, a new automatic rotation recovery algorithm is proposed. This algorithm can be plugged to any image or video watermarking algorithm extraction component. The main job for this method is to detect the geometrical distortion happens to the watermarked image/images sequence; recover the distorted scene to its original state in a blind and automatic way and then send it to be used by the extraction procedure. The work is limited to have a recovery process to zero padded rotations for now, cropped images after rotation is left as future work. The proposed algorithm is tested on top of extraction component. Both recovery accuracy and the extracted watermarks accuracy showed high performance level.**

***Keywords*—Rotation recovery; image watermarking; video watermarking; watermark extraction; robustness**

I. Introduction

Information security and privacy issues have occupied a huge area in the field of computer related research due to the fabulous evolution in information and data exchange. Currently, there is a fast emerge of various methods that allow parties to exchange the media files, starting from social media websites to images and videos sharing utilities and ending in mobile applications such as Whatsapp, Viber, Wechat and many more [1] . This has attracted researchers to increase the focus on securing, authenticating and protecting the exchanged data from malicious attackers. One of the protection techniques that researches were noticed to be focusing on recently is digital watermarking. This focus has led to emerge of various innovations in the image and video watermarking with different algorithms and techniques. In Digital watermarking, the media file is protected by inserting a code called 'watermark', which can be text, image or binary stream, into the host file, which can be an image or video [2]. This watermark will be used for many purposes such as, authentication, copyright protection, forgery detection, leaking protection; where it can be extracted later from the protected media file.

Although, there are different algorithms in digital watermarking for images and videos, and many of them claim high performance in term of robustness, it is still difficult for developers to come up with a perfect algorithm that survives all attacks at once and with high extraction accuracy results. This is due to the tries to achieve a performance tradeoff between various metrics such as imperceptibility and robustness [3]. One of the problems while designing a digital watermarking algorithm is losing the robustness for some attacks, such noising and compression once you concentrate to increase the robustness for geometrical attacks like rotation, scaling, and translation.

The focus here will be on rotation attack by proposing a different solution that adds a new facility to watermarking systems. By plugging the proposed solution to any watermarking system, there will be no need to focus on the design phase whether the developed watermarking algorithm has to be invariant to the rotation attack or not. The proposed algorithm is implemented to be used on top of extraction function. Hence, the algorithm will recover the attacked rotated image or a sequence of images from a video then send the restored images to be used in the extraction phase.

The proposed algorithm scope can be further extended to be integrated to various practical applications other than digital watermarking, such as 3D modeling, image visual enhancement, scene recovery in cameras, and more.

The rest of the paper is organized as follows: the second section reviews the recent and related works. The third section presents the proposed algorithm. The fourth section illustrates the experiments. The fifth section presents the results of evaluating the proposed rotation recovery algorithm. Finally, the sixth section concludes the paper.

II. Related Works

In digital watermarking systems, algorithms to watermark images and videos need to address various performance metrics such as imperceptibility, robustness and capacity. Most researchers try in their algorithms to achieve some tradeoff between imperceptibility and robustness. This tradeoff makes developers to sacrifice some robustness values. For example, increasing the tolerance to some noising attacks with decreasing the visual effects in the watermarked image might lead to losing the resistance to geometrical attacks such as scaling, rotation and translation. Focusing on resisting geometrical attacks might force the developer to sacrifice the visual image quality. For these reasons, there were algorithms that focused to achieve specific tolerance to determined attacks; some examples of these algorithms that intended to be invariant to geometrical attacks are in [4, 5, 6, 7].

In addition, current researches in image watermarking as well as in video watermarking have shown low robustness in the case of rotation attacks. Some examples of low resistance to rotation attacks are obviously reported in [8, 9, 10, 11, 12, 13, 14]. In these works the watermark after rotation attack was difficult to be accurately extracted and the data was mostly lost. The reported normalized correlation (NC) values were very low. Table 1 summarizes some of the reported result in the case of rotation attacks.

TABLE I. WEAK ALGORITHMS UNDER ROTATION ATTACKS

Algorithm	NC	BER
Rakesh Ahuja et al [8]	0.73	-
L. Agilandeeswari et al [9]	0.80	9.33
Ta Minh Thanh et al [10]		-
Nasrin M. Makbol [11]	-	0.50
Zhao et al [12]	0.86	-
Jiansheng et al [13]	0.50	-
Lusson et al [14]	0.65	-

Some approaches were proposed to recover images into their original states after rotation attacks occur. In [15] they proposed a rotation estimation and recovery algorithm using image alignment, Radial Tchebichef moments or Fourier descriptors. In their algorithm, for all used methods, they need the original non-rotated image to be used as a reference image for the recovery and estimation purposes. Although the algorithm works with presence of the reference non-rotated image, it was still estimating the rotation angle with degree error reached to 4 degrees.

Another algorithm previously was proposed by Morgan McGuire [16]. In this algorithm, image registration using Fourier-Mellin transform was used for the purpose of estimating geometrical attacks parameters such as scaling, rotation and translation parameters. The algorithm showed good performance in rotation recovery. However, it reported an error reached up to 1 degree. Moreover, experiments were reported for rotated scenes and not with zero padded images. It also needs a reference image to estimate the parameters.

Previously, a symmetric reversible method was proposed by Laurent Condat and Dimitri Van De Ville [17]. In this method, a 1-D filter is designed to convolve the rotated image with appropriate fractional delay filters. Pixels interpolation was utilized to recover the rotated scene. However, their results showed blurred images after recovery. This indicates the weakness in the algorithm to perfectly recover the rotations. Such algorithm cannot work accurately while used with digital watermarking systems.

In 3D images field, some algorithms have been released to deal with rotated scene estimation. In [18] authors developed a method to automatically recover image rotations from 3D urban scene. This was achieved by estimating various parameters from the taken images by multiple cameras. The parameters such as, intrinsic camera parameters and extrinsic pose are used with edge detection algorithm and vanishing points to estimate the rotation of the scene. This algorithm seems impractical to work with single 2D image rotation attacks, for example, in digital watermarking applications.

Another work in [19] introduced a recovery algorithm for rotations on 3D cameras. The algorithm was developed as a part of creating view panoramic mosaics scene. The algorithm works by registering a sequence of images after recovery rotation and using the registered images to estimate the focal length. However, this algorithm cannot serve some applications such as digital watermarking while it needs to multi mages to recover rotation parameters.

As noticed from Table 1, numerous algorithms have been published claiming high robustness. That is true when considering the common attacks and ignoring geometrical attacks. However, investigating these algorithms proved that rotation attacks still uncovered when algorithms can tolerate other attacks such as noising, filtering, cropping and compression. Low NC and high BER values, under rotation attacks for the investigated algorithms open the way for researchers to find an alternative solution that does not affect the other results while considering different attacks to achieve high performance.

For the purpose of solving the issue of the low performance related to rotation attacks, an alternative solution that adapts the rotation recovery scenario to watermarking systems is proposed instead of designing the watermarking algorithms to be invariant to rotation attacks but affecting other performance metrics.

III. PROPOSED ALGORITHM

As discussed in the previous sections, the main problem was found in the weakness of the most available algorithms for image and video watermarking to resist rotation attack. In consequence, the detection process of the embedded watermarks will be inaccurate if not impossible.

To solve this problem a new recovery rotation algorithm is proposed here to prepare the attacked image before performing extraction process. The algorithm is implemented to be pluggable to the extraction component in any image or watermarking system. This algorithm is developed to automatically detect, estimate and recover rotations for acute angles rotated scenes without need to a reference image. This is due to the difficulty to estimate the acute rotations in images accurately. In addition to the large distortion happens to watermark data once the scene is rotated with acute angles.

To fully recover the rotated image into its original state, the algorithm is implemented to take the attacked image as input, detect the edges in the image, estimate and compute the rotation angle, estimate the original image size then according to the estimated angle and size the rotation recovery is performed. Fig 1. describes the proposed algorithm process.

A. Recovery algorithm

The proposed rotation recovery algorithm is implemented in the following steps. Figure 2 shows the states of the image during executing the recovery algorithm.

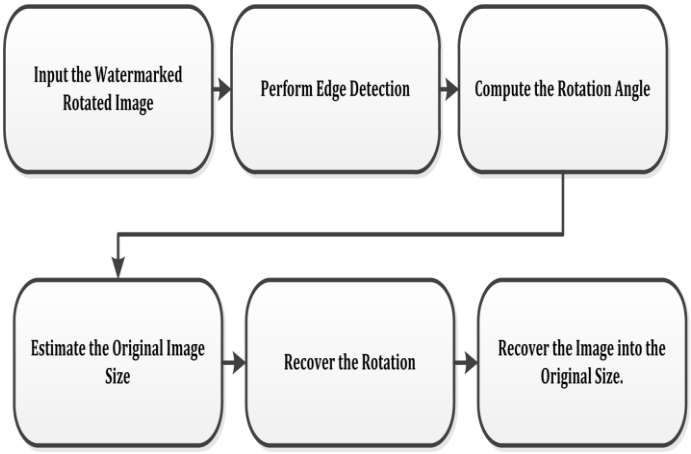


Fig. 1. Rotation Recovery Process

Step 1: Input the rotated image *I*.

Step 2: if *I* is in RGB, Convert *I* to gray scale image.

Step 3: Apply *Canny Edge Detector* to detect edges in image *I*.

Step 4: Apply image dilation on the output from Step 3 using desk structuring element of radius = 3. Save the result as *EdgeImage*.

Step 5: Measure the rotation angle as follow:

- Set one flag carrying first pixel value from the image coroner such that:

$$\text{Flag} = \text{EdgeImage}(1,1) \quad (1)$$

- Measure the opposite side of the angle: Loop in image rows and count the pixel values such that:

$$\text{Opposite} = \begin{cases} \text{Opposite} + 1, & \text{EdgeImage}(\text{row}, 1) = \text{Falg} \\ \text{Opposite} \quad (\text{Break loop}), & \text{Otherwise} \end{cases} \quad (2)$$

Where *Opposite* is set to '0' the beginning.

- Measure the adjacent side of the angle: Loop in image columns and count the pixel values such that:

$$\text{Adjacent} = \begin{cases} \text{Adjacent} + 1, & \text{EdgeImage}(1, \text{col}) = \text{Falg} \\ \text{Adjacent} \quad (\text{Break loop}), & \text{Otherwise} \end{cases} \quad (3)$$

Where *Adjacent* is set to '0' the beginning.

- Calculate the angle using opposite and adjacent lengths [20] where:

$$\text{Angle} = \text{round}(\tan^{-1} \frac{\text{Opposite}}{\text{Adjacent}}) \quad (4).$$

Step 6: Rotate the image *I* by *Angle* and save the rotated image as *RImag*e where:

$$\text{Angle} = -1 * \text{Angle} \quad (5)$$

Step 7: Estimate the original image size using *RImage* as follow:

- Set two flags from *RImage* such that:

$$\text{FlagL} = \text{RImage}(1, \text{round}(\tfrac{\text{L}}{2})) \quad (6)$$

$$\text{FlagH} = \text{RImage}(\text{round}\left(\tfrac{\text{H}}{2}\right), 1) \quad (7)$$

Where *L* is the length of *RImage*, and *H* is the height of *RImage*.

- Loop to calculate the distances BL (Black Length) and BH (Black Height) between the edge of the image and the original scene such that:

$$\text{BL} = \begin{cases} \text{BL} + 1, & \text{RImage}(\text{round}\left(\tfrac{\text{L}}{2}\right), \text{Col}) = \text{FlagL} \\ \text{BL} \quad (\text{Break loop}), & \text{Otherwise} \end{cases} \quad (8)$$

$$\text{BH} = \begin{cases} \text{BH} + 1, & \text{RImage}(\text{row}, \text{round}\left(\tfrac{\text{L}}{2}\right)) = \text{FlagH} \\ \text{BH} \quad (\text{Break loop}), & \text{Otherwise} \end{cases} \quad (9)$$

Where BH and BL are set to '0' in the beginning.

- Find the original image size such that

$$\text{OriginalL} = \text{L} - ((\text{BL} + 1) * 2) \quad (10)$$

$$\text{OriginalH} = \text{H} - ((\text{BH} + 1) * 2) \quad (11)$$

Where *L* and *H* are the length and Height of *RImage.*

Step 8: Crop *RImage* from Point (BL, BH) and by size of *OriginalL* and *OriginalH*.

Step 9: Return the Recovered Image.

B. Edge detection and angle estimation

After converting the colored attacked image to gray scale as explained in section 3.1, and to simplify the estimation of the angle with more accurate value, the edges of the rotated scene are detected using canny operator. The output from canny edge detector will be a binary image (black and white). Although canny operator works perfectly for edges, the resulted image still sometimes can cause inaccurate angle estimation. That is due to the black holes spotted in some places of the edge. These black holes can lead to wrong angle sides' measurements. To solve this matter, an image dilation using desk structuring element of radius = 3 is considered. This will fill the holes in the edge and the angle sides' measurements will be perfectly accurate.

In the dilated image, to estimate the rotation angle, a flag that carries the value of the pixel chosen from the opposite side or the adjacent side can be used to count the similar pixels that have the same value. In current case the pixel value will be '0'. The counting continues until it finds a different pixel value. The stop pixel value is '1'. The count of pixels either for adjacent or opposite side is registered as the length. The length of opposite and the length of the adjacent are used to estimate the rotation angle according to equations 1 to 5. Fig. 3. describes the required measurements for the proposed recovery algorithm.

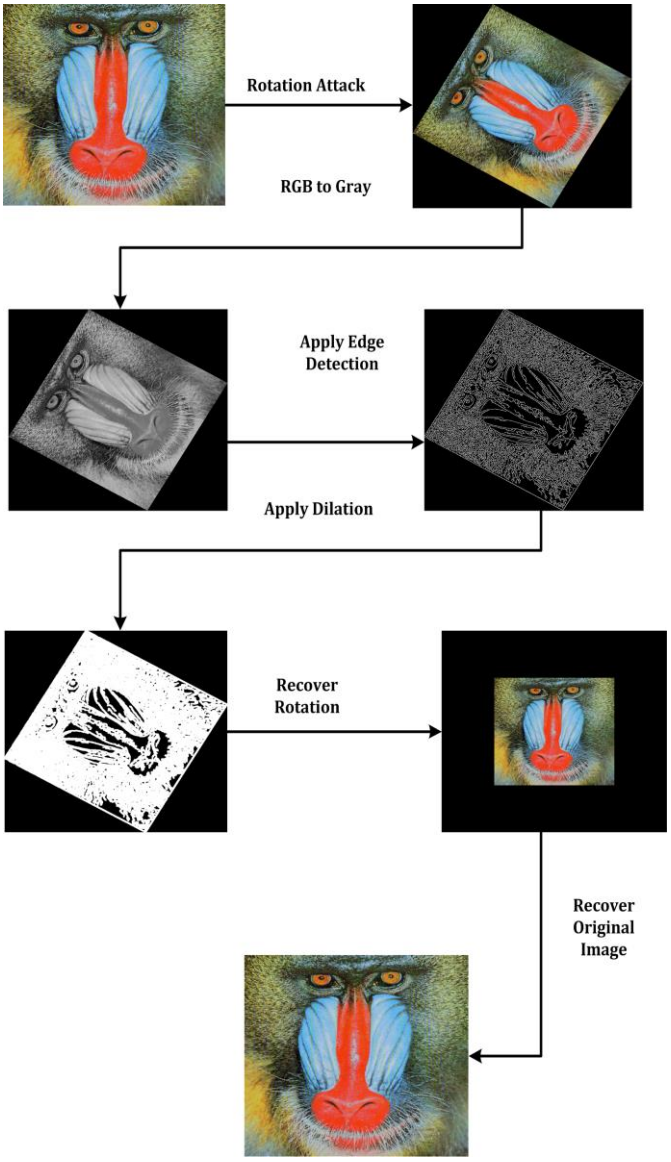


Fig. 2. Image States during Rotation Recovery

C. Image size estimation

Once the rotation angle is estimated and the scene is re-rotated, a black or unwanted area will be resulted with a larger image size than it was in the original one. Two steps must be performed, estimating the original image size form the rotated scene and eliminating the unwanted area. Otherwise, the detection of the watermark data will not be possible.

In the proposed algorithm, the estimation of the size is implemented in a blind manner assuming that the original image size is unknown. To perform this, the distances between the scene and the image edge BH and BL, the original image length (width), (L), the original image height (H) must be measured according to equations 6 to 11. These measurements are then used to crop the recovered scene and eliminate the unwanted area as explored in Fig. 2 and Fig. 3. In this point, the image is ready to be used by the extraction function of the watermarking algorithm which makes the extraction function retrieves the embedded data accurately.

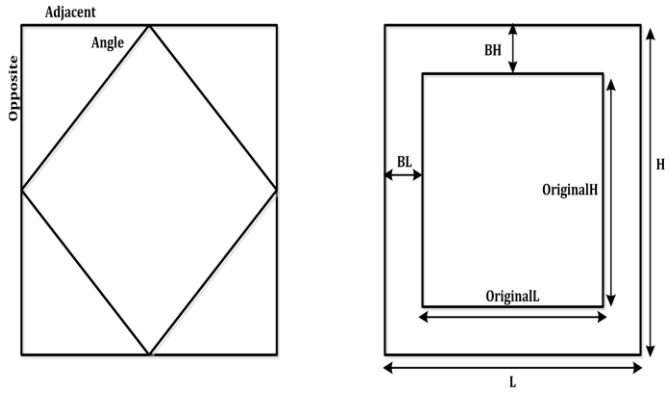


Fig. 3. Required Measurements for Rotation Recovery

IV. Experiments

For the purpose of evaluating the proposed rotation recovery algorithm under digital watermarking environment, the algorithm must be implemented on top of extraction function for any available image watermarking. This is to ensure that the algorithm has attained its main objective, which is to enable digital watermarking systems to resist rotation attacks and increase the accuracy of the extracted watermarks.

To validate the performance of the proposed algorithm, implementing an image digital watermarking algorithm according to [14] is considered. This algorithm was chosen due to its weakness to withstand rotation attacks. This issue made it possible to implement the proposed rotation recovery algorithm on top of the implemented image watermarking algorithm to verify the performance and see how it is possible to survive rotation attack after using the proposed recovery algorithm. The utilized image watermarking algorithm, is implemented based on Discrete Wavelet Transform (DWT), and was developed according to the framework as shown in Fig. 4.

After implementing the image watermarking algorithm based on Fig. 4., and adapting the recovery algorithm to the extraction component, the testing is performed by comparing the extracted watermarks from the image watermarking algorithm before and after using the recovery algorithm. Fig. 5. explains the testing scenario used to evaluate the proposed recovery algorithm.

To emphasis the results, three different standard images Baboon, Lena and Peppers of size 512x512 are used. Each image was watermarked using the implemented image watermarking algorithm, with a watermark of size 64x64 pixels. The watermarked images are then attacked by rotating the images using various acute angles. The watermarks are extracted using both implemented watermarking algorithms with and without recovery.

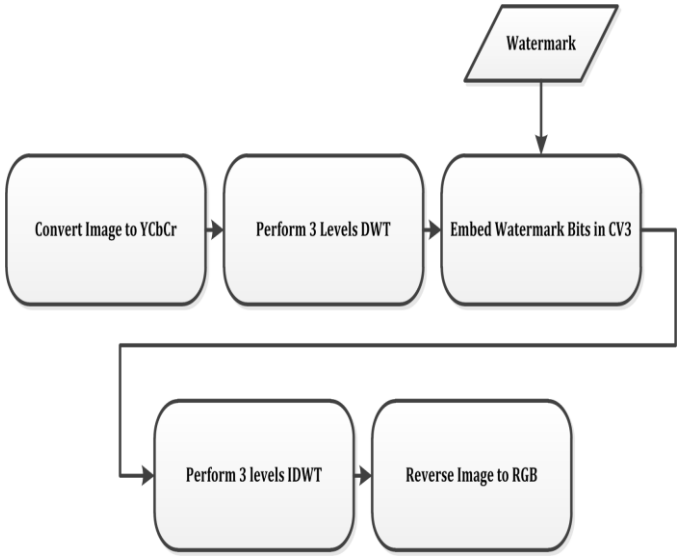


Fig. 4. Image Watermarking Algorithm Framework

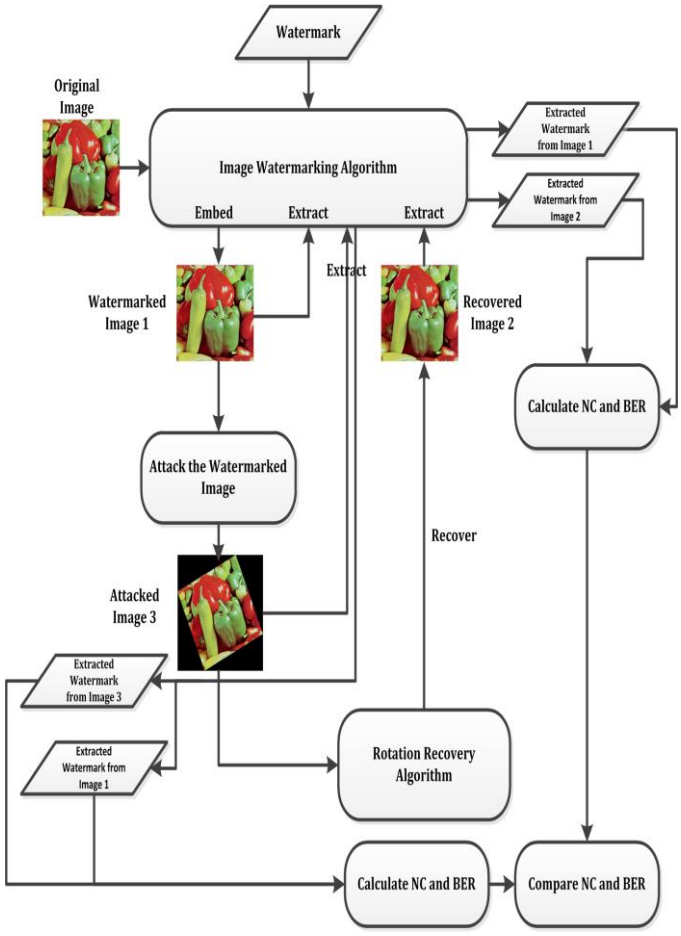


Fig. 5. Testing Scenario

V. Results and Discussion

The testing of the recovery algorithm was conducted using the experiments illustrated in the previous section. Two measures were used to evaluate the accuracy of the extracted watermarks before and after using the proposed rotation recovery algorithm, Normalized Correlation (NC) and Bit Error Rate (BER) according to the following formulas:

$$NC = \left(\frac{\sum_{i=0}^{m-1}\sum_{j=0}^{n-1} O(i,j) - E(i,j)}{\sqrt{\sum_{i=0}^{m-1}\sum_{j=0}^{n-1} O^2(i,j)} \times \sqrt{\sum_{i=0}^{m-1}\sum_{j=0}^{n-1} E^2(i,j)}} \right) \quad (12)$$

Where, O is the original watermark, E is the extracted watermark.

$$BER = \frac{\sum_{i=0}^{m-1}\sum_{j=0}^{n-1} O(i,j) - E(i,j)}{mxn} \quad (13)$$

Where, *O* is the original watermark, *E* is the extracted watermark, and *mxn* is the total number of watermark bits.

After implementing the image watermarking algorithm and testing the detection of the watermarks in the normal case where no rotation attack has applied to the watermarked images, the watermarks were extracted accurately as in Fig. 6.

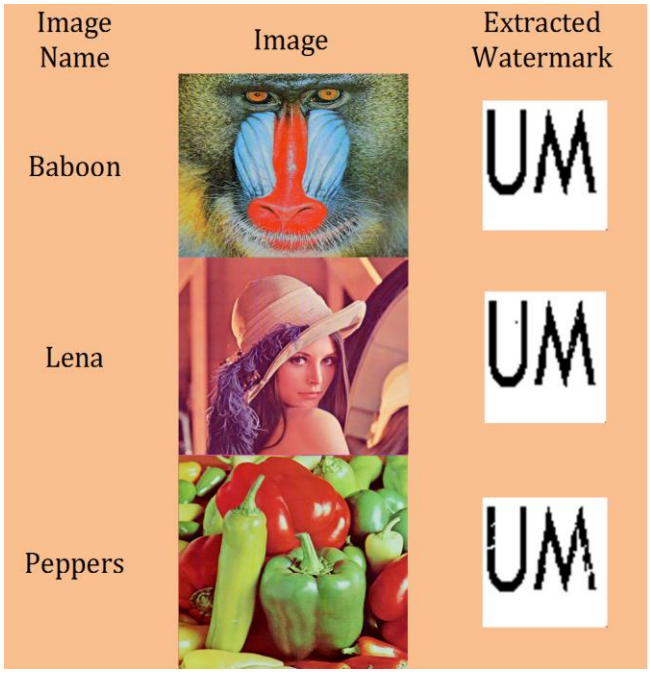


Fig. 6. Extracted Watermarks before Rotation Attacks

In the current case, the extracted watermarks in the situation where no rotation is applied as the original watermarks to be used for calculating NC and BER values for the extracted watermarks after recovery is considered. The comparison is done later with the extracted watermarks after rotation attack.

To evaluate the accuracy of the rotation recovery algorithm, the normalized correlation was measured for both original image after watermarking and the recovered image from rotation attacks. This is to indicate how accurate the recovery is. Results from the used images conducted based on various angles have shown NC of '1' for Lena and Peppers and 0.98 for Baboon image. Baboon image has not reached to NC of '1'

because the error expected on size estimation on some images which is around 1 pixel height or width. However, in recovered images having value of 0.98 of NC, it was enough to retrieve the watermark accurately. Fig. 7. shows the NC values for the various angles rotation recovery in the three images.

After ensuring the high performance of the rotation recovery algorithm, the experiments under the implemented watermarking algorithm was conducted. The mentioned three tested images, Baboon, Lena and Peppers were tested. The extracted watermarks from attacked images with different rotation angles and from recovered images were utilized to measure both NC and BER. Fig. 8., Fig. 9., Fig. 10., Fig. 11., Fig. 12. and Fig. 13. show the results obtained for these experiments.

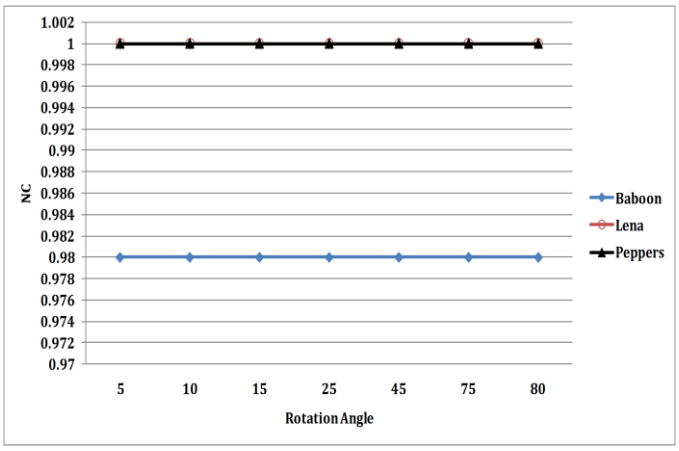


Fig. 7. Rotation Recovery Accuracy

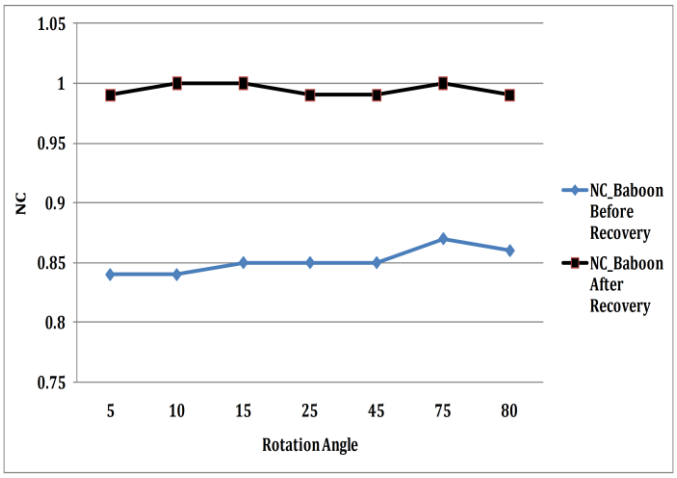


Fig. 8. Normalized Correlation for Extracted Watermarks for Baboon Image

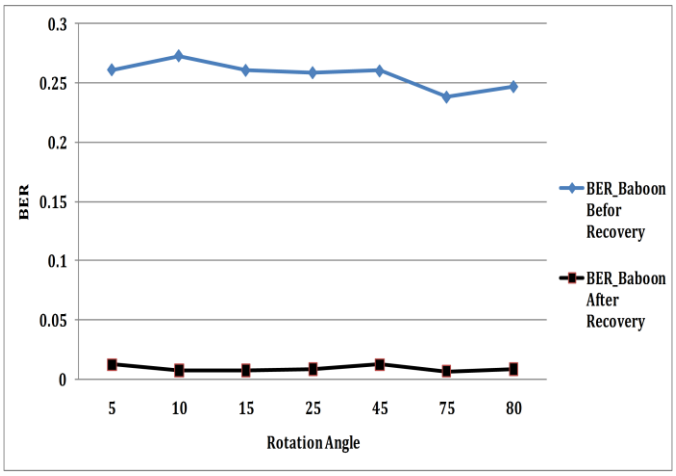


Fig. 9. Bit Error Rate for Extracted Watermarks for Baboon Image

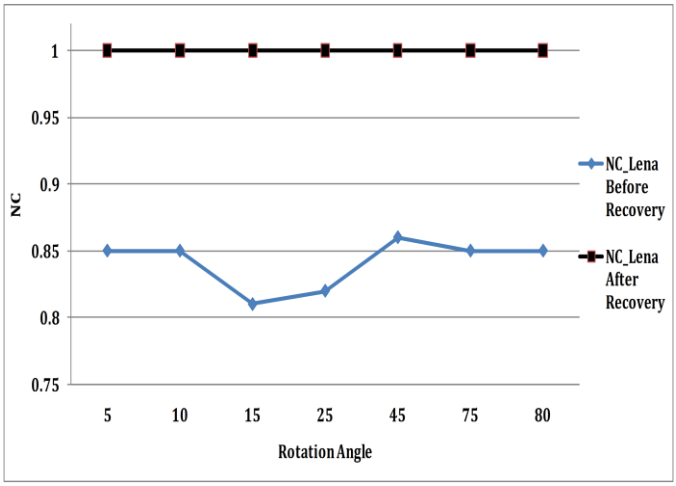


Fig. 10. Normalized Correlation for Extracted Watermarks for Lena Image

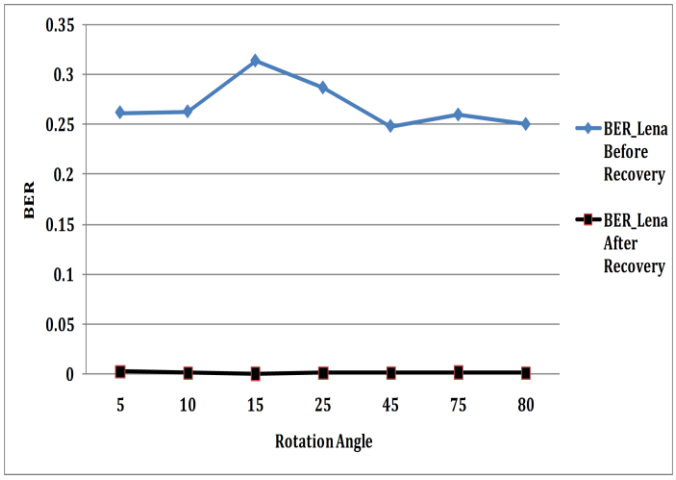


Fig. 11. Bit Error Rate for Extracted Watermarks for Baboon Image

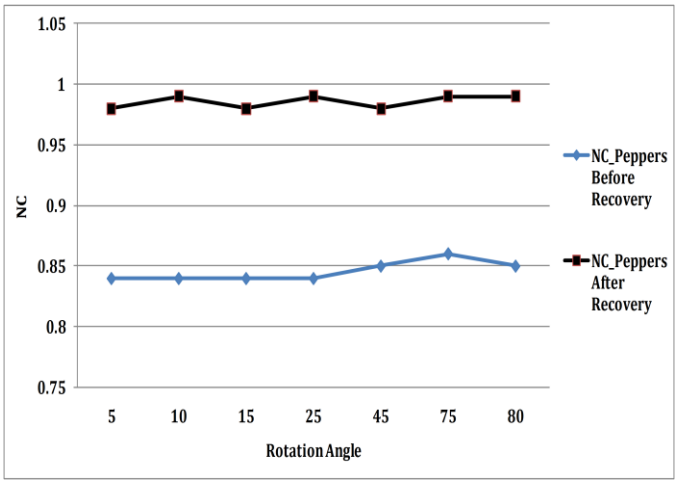


Fig. 12. Normalized Correlation for Extracted Watermarks for Peppers Image

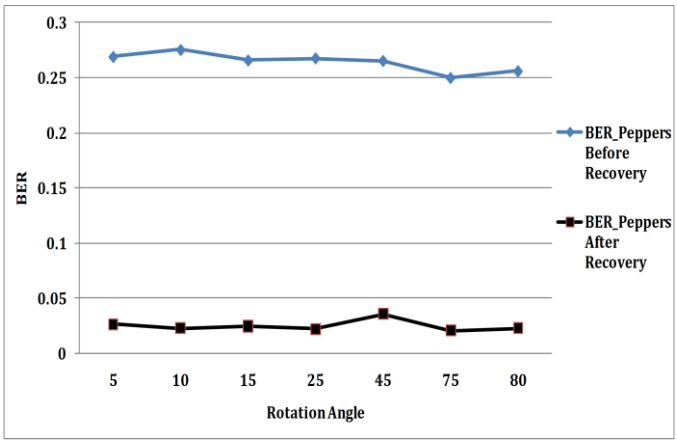


Fig. 13. Bit Error Rate for Extracted Watermarks for Peppers Image

As seen in the results above, the NC value was extremely improved in the extracted watermarks from around 0.80 in the attacked images to 1 in the recovered images in most cases. At the same time BER value improvements from around 0.30 to almost 0.000 was achieved in most cases. These results prove that the proposed rotation recovery algorithm was developed to perfectly suits image watermarking systems and increases the resistance against rotation attacks.

As shown in Fig. 14. a sample of the extracted watermarks from attacked and recovered images is presented. The extracted watermarks are almost lost with attacks applied to the watermarked images. In contrast, the watermarks were accurately extracted after performing the recovery using the proposed algorithm.

VI. Conclusion

In this paper, the current digital image and video watermarking algorithms were investigated in term of robustness. Consequently, the weakness of the most algorithms was noticed to be resides on surviving the rotation attacks. This article has proposed a new automatic and blind algorithm to recover acute angles rotations in images. The proposed algorithm estimates the angle of rotation mathematically then estimates the original watermarked image size in a blind way. Based on the angle and estimated size, the original image is recovered. The proposed algorithm has been made pluggable to the extraction function in any watermarking algorithm to benefit from extracting accurate watermarks. Using such algorithm will save developers from considering rotation attacks during the design of the watermarking algorithms. This work has been designated for zero padded rotated images while investigating images by applying cropping attack after rotation attack is left as an improvement for the current work on the future. Testing the proposed algorithm showed NC of '1' for the recovery process which indicates accurate angles estimations. Evaluating the adaption under digital watermarking algorithms showed very high accuracy for the extracted watermarks as well. Regardless of specifically using the proposed rotation recovery algorithm for digital watermarking, it can be proudly integrated to other image processing applications.

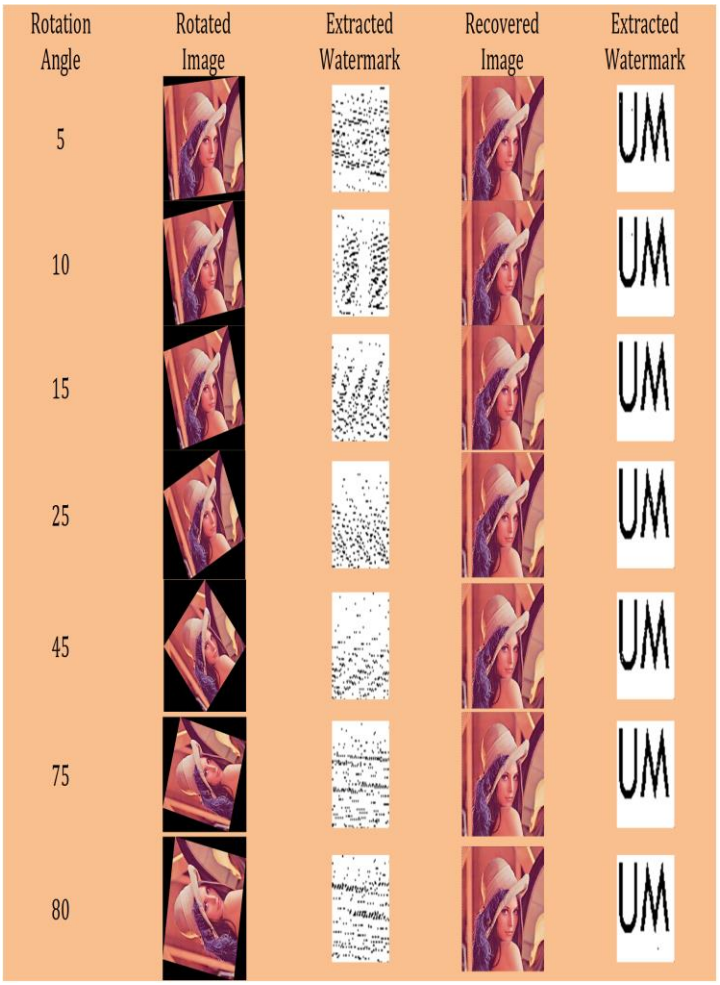


Fig. 14. Extracted watermarks, Attacked vs. Recovered Images

Acknowledgments

This work is funded under the Fundamental Research Grant Scheme; grant number FP061-2014A awarded by the Ministry of Higher Education, Malaysia, for the period of July, 2014 until end of June, 2016 and the University of Malaya's Research Grant (UMRG), grant number RP030A-14AET.

REFERENCES

[1] J. C. Bertot, P. T. Jaeger, and D. Hansen, "The impact of polices on government social media usage: Issues, challenges, and recommendations," *Government Information Quarterly*, vol. 29, no. 1, pp. 30 – 40, 2012. [Online]. Available: http://www.sciencedirect.com/-science/article/pii/S0740624X11000992

[2] I. Cox, M. Miller, J. Bloom, J. Fridrich, and T. Kalker, *Digital watermarking and steganography*. Morgan Kaufmann, 2007.

[3] S. Stankovic, I. Orovic, and E. Sejdic, *Multimedia Signals and Systems: Basic and Advanced Algorithms for Signal Processing*. Cham: Springer International Publishing, 2016, ch. Digital Watermarking, pp. 349–378. [Online]. Available: http://dx.doi.org/10.1007/978-3-319-23950-7_7

[4] W. Chun-peng, W. Xing-yuan, and X. Zhi-qiu, "Geometrically invariant image watermarking based on fast radial harmonic fourier moments," *Signal Processing: Image Communication*, vol. 45, pp. 10 – 23, 2016. [Online]. Available: http://www.sciencedirect.com/science/article/pii/-S0923596516300236

[5] S. Fazli and M. Moeini, "A robust image watermarking method based on dwt, dct, and {SVD} using a new technique for correction of main geometric attacks," *Optik - International Journal for Light and Electron Optics*, vol. 127, no. 2, pp. 964 – 972, 2016. [Online]. Available: http://-www.sciencedirect.com/science/article/pii/S0030402615012863

[6] R. Koju and S. R. Joshi, "Performance evaluation of slant transform based gray image watermarking against common geometric attacks," *International Journal of Computer Science and Information Security*, vol. 14, pp. 137–150, 2016.

[7] X.-j. Wang and W. Tan, *Proceedings of the 6th International Asia Conference on Industrial Engineering and Management Innovation: Innovation and Practice of Industrial Engineering and Management (volume 2)*. Paris: Atlantis Press, 2016, ch. An Improved Geometrical Attack Robust Digital Watermarking Algorithm Based on SIFT, pp. 209–217. [Online]. Available: http://dx.doi.org/10.2991/978-94-6239-145-1_21

[8] L. Agilandeeswari and K. Ganesan, "A robust color video watermarking scheme based on hybrid embedding techniques," *Multimedia Tools and Applications, Springer*, 2015.

[9] R. Ahuja and S. S. Bedi, "Copyright protection using blind video watermarking algorithm based on mpeg-2 structure," in *International Conference on Computing, Communication and Automation (ICCCA2015)*. IEEE, 2015, pp. 1048–1053.

[10] F. Lusson, K. Bailey, M. Leeney, and K. Curran, "A novel approach to digital watermarking, exploiting colour spaces," *Signal Processing*, vol. 93, no. 5, pp. 1268 – 1294, 2013. [Online]. Available: http://-www.sciencedirect.com/science/article/pii/S0165168412003805

[11] N. M. Makbol, B. E. Khoo, and T. H. Rassem, "Block-based discrete wavelet transformsingular value decomposition image watermarking scheme using human visual system characteristics," *IET Image Processing*, vol. 10, pp. 34–52, 2016.

[12] Mei Jiansheng, Li Sukang and Tan Xiaomei, "A Digital Watermarking Algorithm Based On DCT and DWT," in *Proceedings of the 2009 International Symposium on Web Information Systems and Applications (WISA 09)*, Nanchang, P. R. China, May 2009, pp. 104–107.

[13] T. M. Thanh, P. T. Hiep, T. M. Tam, and K. Tanaka, "Robust semi-blind video watermarking based on frame-patchmatching," *International Journal of Electronics and Communications (AEÜ),Elsevier*, vol. 68, pp. 1007–1015, 2014.

[14] Yanxia Zhao and Zenghui Zhou, "Multipurpose Blind Watermarking Algorithm for Color Image Based on DWT and DCT," in *2012 8th International Conference on Wireless Communications, Networking and Mobile Computing (WiCOM)*,, Sept 2012, pp. 1–4.

[15] S. M. Elshoura and D. B. Megherbi, "International symposium on computer, communication, control and automation," in *International Symposium on Computer, Communication, Control and Automation*. USA: IEEE, 2010.

[16] M. McGuire, "An image registration technique for recovering rotation, scale and translation parameters," Technical Report, 98-018, NEC Research Institute, Tech. Rep., 1998.

[17] L. Condat and D. V. D. Ville, "Fully reversible image rotation by 1-d filtering," in *IEEE International Conference on Image Processing*, vol. 8. IEEE, 2008, pp. 913–916.

[18] Antone, M. E. & Teller, S. "Automatic recovery of relative camera rotations for urban scenes", *Computer Vision and Pattern Recognition, 2000. Proceedings. IEEE Conference on*, **2000**, *2*, 282-289 vol.2

[19] Szeliski, R. & Shum, H.-Y., "Creating Full View Panoramic Image Mosaics and Environment Maps", Proceedings of the 24th Annual Conference on Computer Graphics and Interactive Techniques, ACM Press/Addison-Wesley Publishing Co., **1997**, 251-258.

[20] Tim Alderson, "Calculus of Trigonometric Functions," (Department of Mathematics, The University of Western Ontario, Canada), 2003. Lecture Notes.

A Comparative Study Between the Capabilities of MySQl Vs. MongoDB as a Back-End for an Online Platform

Cornelia Győrödi
Department of Computer Science and Information Technology, University of Oradea
Oradea, Romania

Ioana Andrada Olah
Department of Computer Science and Information Technology, University of Oradea
Oradea, Romania

Robert Győrödi
Department of Computer Science and Information Technology, University of Oradea
Oradea, Romania

Livia Bandici
Faculty of Electrical Engineering and Information Technology, University of Oradea
Oradea, Romania

***Abstract*—In this article we present a comparative study between the usage capabilities of MongoDB, a non-relational database, and MySQL's usage capabilities, a relational database, as a back-end for an online platform. We will also present the advantages of using a non-relational database, namely MongoDB, compared to a relational database, namely MySQL, integrated in an online platform, which allows users to publish different articles, books, magazines and so on, and also gives them the possibility to share online their items with other people. Nowadays, most applications have thousands of users that perform operations simultaneously thus, it takes more than one operation to be executed at a time, to really see the differences between the two databases. This paper aims to highlight the differences between MySQL and MongoDB, integrated in an online platform, when various operations were executed in parallel by many users.**

***Keywords*—MySQL; relational database; MongoDB; non-relational database; comparative study**

I. Introduction

Nowadays, an application must be accessible to its users 24 hours a day, 7 days a week, so it is important to implement an appropriate database, which supports simultaneous connection of hundreds of thousands of users [6]. Also, more and more complex requirements from users appeared, and companies were forced to find different solutions to meet the needs of their customers. Thus, the applications must support millions of users simultaneously and handle a huge volume of data and a relational database model has serious limitations when it has to handle that huge volume of data. Because each customer has his own needs and requirements, it might be possible that within the same application, there is a need for a different customization for each user. Relational databases do not allow a complete configuration that can shape after their needs. These limitations have led to the development of non-relational databases, also commonly known as NoSQL (Not Only SQL) [11]. The NoSQL term was coined by Carlo Strozzi in 1998, and refers to non-relational databases, term which was later reintroduced in 2009 by Eric Evans [2, 6].

Non-relational databases do not use the RDBMS principles (Relational Data Base Management System) and do not store data in tables, schema is not fixed and have very simple data model [6]. Their main advantage is represented by their flexible structure, but also because they are designed in a way that can store large amount of data. In addition, they are denormalized databases, which leads to increased performance [1].

In this paper, we will try to present the advantages of using MongoDB compared to MySQL, integrated in an online platform, which allows users to publish different articles, books, magazines and so on, and gives them the possibility to share online their items with other people. At the same time, we will show a comparison between the two databases, MongoDB and MySQL, in terms of execution times when many users executed various operations in parallel.

II. Presentation of the Application Database

Databases to be presented are part of an online application that allows its users to create interactive digital flipbooks. Basically, the application could be called an online library, where users can create digital books that can be accessed by users worldwide.

Such an application can be used by many types of users, such as writers (who want to submit previews of their books), teachers (who can publish educational tutorials for their students), companies or supermarkets (that wish to advertise online or to submit weekly brochures) and so on.

NoSQL databases provide you with ways of storing and retrieving the data that is not modelled as the relational databases are modelled. Mainly, NoSQL databases are designed to allow us insertion of data for which we do not have a predefined schema, as the structure of our data is not set [7].

This work was performed through the Partnerships Program in priority areas, PN-II-PT-PCCA-2013-4-2225 - No. 170/2014 developed with the support of MEN - UEFISCDI, "Electromagnetic methods to improve processes wine".

A database like MongoDB does not a have the concept of a "row"; instead, we have a more flexible model [4]. There are four strategies for storing data in a non-relational database, as shown in [5]. They are not based on a single model (e.g. relational model of RDBMSs) and each database, depending on their target-functionality, adopt a different one [9]. The design of NoSQL databases depends on the type of database, called store. Document Stores pair each key identifier with a document which can be a document, key-value pairs, or key-value arrays. Graph Stores are designed to hold data best represented by graphs, interconnected data with an unknown number of relations between the data [10].

As a MongoDB database does not actually have a graphical representation, the tabular representation for the MySQL database will be presented. The two databases have the same number of columns and the same data type for each field. At a structural level, the only difference between the two databases is the way their data is represented.

The database can be divided into 3 parts, namely:

- users' area;
- items' area;
- orders' area.

A. *Users' area* – lists tables containing information about all the users, such as address, country, city or phone number as shown in Fig. 1.

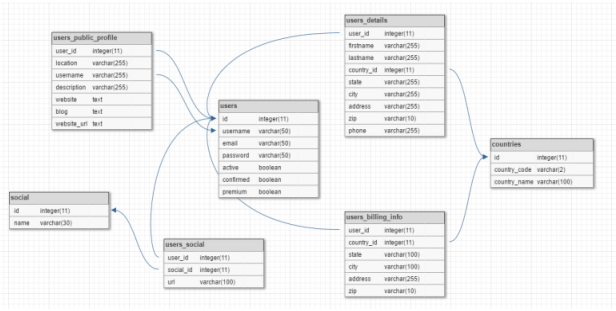


Fig. 1. Users' tables

As reflected in the figure Fig. 1, the users area contains 7 tables, 2 of which are static: *countries* (which contains a list of all existing countries) and *social* (which contains a list of names of social networks, such as Facebook and LinkedIn, which users can add to their public profile, to promote themselves).

The other tables, listed below, contain the following fields:

- *users* table, with the fields *id*, *username*, *email*, *password*, *active*, *confirmed* and *premium* (this table contains the user's data required for registration);
- *users_billing_info* table, with the columns *user_id*, *country_id*, *state*, *city*, *address* and *zip*, which stores the user's billing info data. When a user registers on the platform, it has by default the Free Package. The user may subsequently change the package, buying a new one;
- *users_details* table, with the optional fields *firstname*, *lastname*, *country_id*, *state*, *city*, *address*, *zip*, *phone*, which the user can set or not;
- *users_public_profile*, which contains information about the user's public profile, such as personal website or blog page;
- *users_social*, which stores data about the user's social networks (Facebook, Google+, LinkedIn etc.).

Also, the users' area is closely linked to 2 static tables, which are the *countries* and the *social* tables. The link between the tables is made by foreign keys, which are either the country id, or the social network id.

B. *Items' area* – contains information about the items ploaded by a user (books, magazines, catalogs), as well the SEO categories to which they belong as shown in Fig. 2.

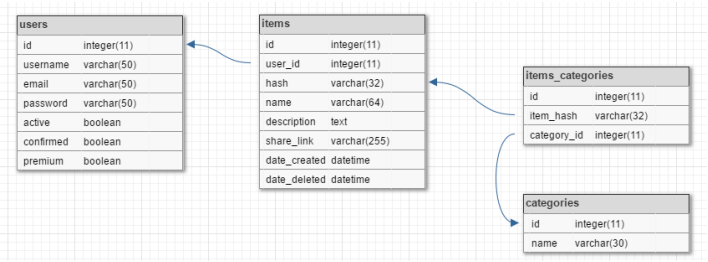


Fig. 2. Items' tables

In the figure Fig. 2 are presented the tables containing data about the items a user that can be upload on the platform.

The *items* table contain explicit data about each item, such as the name and description set by the user, the date on which it was created, respectively the date on which it was deleted (which is *null* by default), a hash (a string representing the surest way of identifying an item from the platform, because it is unique) and a *share link* that is automatically formed from the name of the platform and the item hash.

Also, as the database described in this application can be integrated into a promotional platform, each user has the possibility to add different categories for each item, in order to be found faster by search engines. The categories are static and are found in the *categories* table.

C. *Orders' area* – contains information about the users' rders and transactions (the premium package bought by a user, the amount of the package, the date the order was made and so on) is shown in Fig. 3.

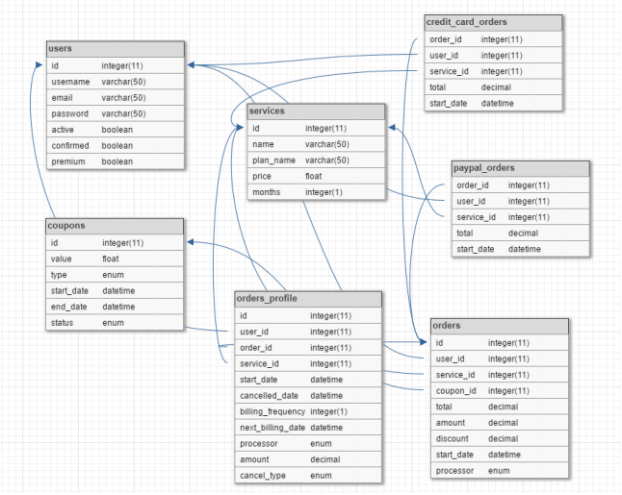


Fig. 3. Orders' tables

The orders area is a rather complex part in any application, because it must communicate with all other parties. In this application, an order refers to a user switching from one package to another (packages will be listed later). Thus, the *orders* table, that effectively represents all the transactions made by users, contains the following fields:

- *id* – a unique field, automatically generated, that identifies the order;
- *user_id* – id of the user who made the order (the user id is also unique);
- *service_id* – the type of the package that the user has bought (the types of packages that exist in an application are varied and they depend greatly on the type of application concerned). For this case study, we used the following packages: User Pro, User Business and User VIP (each package can be purchased either for a period of one month, or for a period of 12 months);
- *coupon_id* – most applications (websites, online platforms, shops and many others) offer various discounts to its users. These discounts are usually provided with discount coupons, which can be applied in most cases only once, and can be of two main categories: numerical coupons and percentage coupons;
- *amount* – the amount of the package chosen by the user, and which is calculated by multiplying the number of months selected with the package price (usually, the annual packages are cheaper than the monthly ones, and offer a discount by default);
- *discount* – the amount of discount that is given (or not) to a user, when they conduct a transaction (when they purchase a package);
- *total* – the total amount that the user has to pay, and which represents the difference between the package and the discount offered;
- *start_date* – the date on which the package was bought;
- *processor* – the payment method chosen by the user (the most commonly used payment methods are PayPal and Credit card).

Also, depending on the payment method chosen by the user, all the data related to the transaction will be inserted either in the *credit_card_orders* table, if the payment was made by credit card, or in the *paypal_orders* table, if payment is made through PayPal.

Another important table is the table *orders_details.* This table contains information about the next date when user will be billed, i.e. the date when the current subscription will expire and when the current package will have to be automatically renewed. In most applications, once a user has purchased a package, it will be renewed automatically, once a month or every 12 months, depending on the type of package that the user owns, but only if the user does not manually cancel the subscription.

III. PERFOMANCE TESTS

To highlight the advantages of using a non-relational database, MongoDB, compared a relational database, MySQL, various operations were performed on the two databases in parallel by many users. These operations represent the four elementary operations that can be performed on any database, namely: insert, select, update and delete [3].

All the tests to be presented were conducted on a computer with the following configuration: Windows 7 Pro 64-bit, processor Intel Core i3 (2.4 GHz), 4 GB RAM memory.

To have data on which to carry out operations, some data had to be inserted. Because in any application, the users are the most important part and without them, the application would not exist, the test started with the insertion of users in both databases (MySQL and Mongo, respectively).

To generate user's data such as username, email address, and password, various PHP functions [8], such as *md5*, *rand*, *substr* and *str_shuffle* were used. In fact, the functions listed above were used to generate all data for the databases, such as city, address, telephone number, personal website, item name or description. In order to record the time required to insert the elements in the database, it was used the PHP function *microtime*, which recorded the time from the beginning of the script runtime and until its completion.

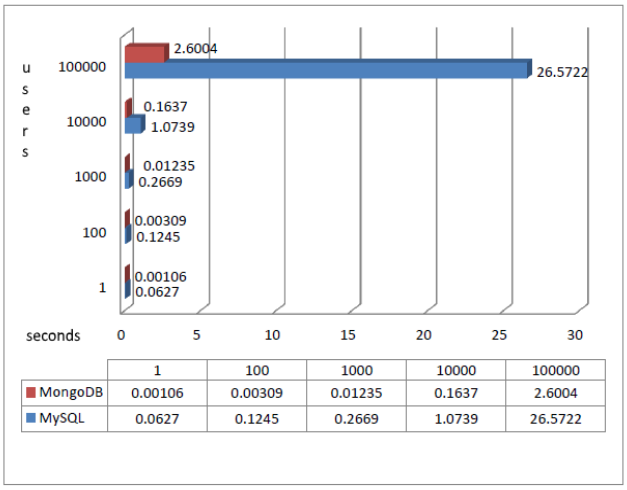
Most applications and websites give users the possibility of creating a public profile, after they successfully register on the site. Since not all users who make an account in an application complete all the necessary data for a public profile, and they only create a simple account, we conducted two types of tests for inserting users.

The first test refers only to a simple user registration on the website, which is the creation of an account only by setting a

username, *email address*, and *password*. These data are listed in the table of *users*.

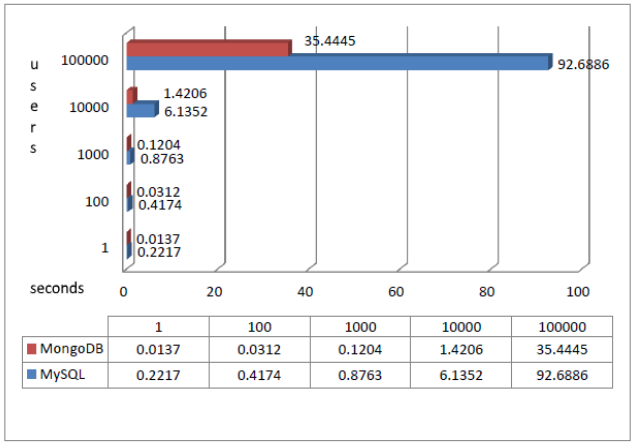
The second test includes the creation of a public profile and the insertion of some additional data, such as country, city, phone number, billing details or links to various social networks. These data are entered, as appropriate and as described in the previous chapter, in the tables *users_billing_info*, *users_details*, *users_public_profile* and *users_social*.

To test the performance in terms of speed differences between the two databases, 5 tests were performed for each insert case (5 tests for users who only register on the platform, and 5 tests for users who also completed some other details). The number of users has varied, starting from 1 user and continuing with 100, 1.000, 10.000 and 100.000 users respectively.



	1	100	1000	10000	100000
MongoDB	0.00106	0.00309	0.01235	0.1637	2.6004
MySQL	0.0627	0.1245	0.2669	1.0739	26.5722

Fig. 4. Insert users without details



	1	100	1000	10000	100000
MongoDB	0.0137	0.0312	0.1204	1.4206	35.4445
MySQL	0.2217	0.4174	0.8763	6.1352	92.6886

Fig. 5. Insert users with details

As is showed in the graphs from Fig. 4 and Fig.5, we can see that MongoDB has a good performance; it was faster than MySQL in all insert cases. However, this is best proven by the insertion of 100.000 users, where MongoDB was 2 times faster when the users only registered on the platform, and 10 times faster, when users also had completed some other personal details.

Following the results from Fig. 4 and Fig. 5, we can be easily seen that MongoDB has a much higher 'working speed' compared to the speed of MySQL. However, although the data obtained above is true and accurate, those tests are not applicable in everyday life. Tests carried out above represent one operation at a time. Although there are small applications and websites that have a relatively small number of users, nowadays, most applications have thousands of users that perform operations simultaneously. In other words, it takes more than one operation at a time, to really see the differences between the two databases.

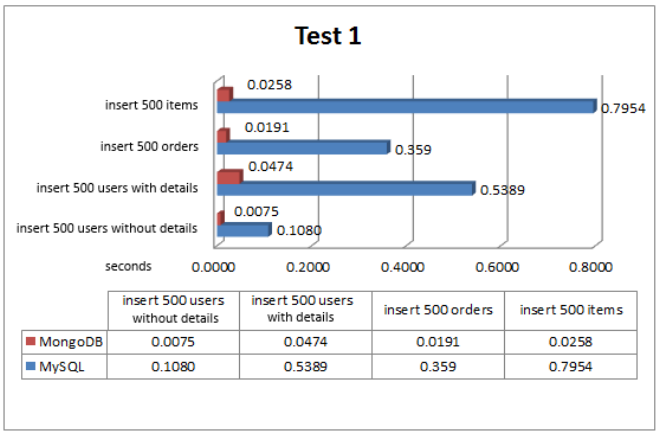
This paper aims to highlight the differences between MySQL and MongoDB, using the databases described above, in a real application, where various operations (such as: data is inserted, data is deleted or data is modified) were executed in parallel by many users.

Thus, we will describe further three tests, each with different difficulty level in order to accurately calculate the differences between the two databases.

A. *Test 1*

For the first test, 1.000 users were previously inserted in both databases. Then, the test actually begins with the insertion of 1.000 users in the databases, of which 500 users only register on the platform, and the remaining 500, also complete other details (described in a previous test). At the same time, there are 500 orders (half of the existing users buy a premium package) and other 500 users create (upload) different items. All tests in this paper were achieved using the PHP programming language, and their execution was carried out in parallel using the PHP *exec* function [8], that executes these operations in parallel.

The results obtained after carrying out the first test described above are illustrated in the figure Fig. 6.



	insert 500 users without details	insert 500 users with details	insert 500 orders	insert 500 items
MongoDB	0.0075	0.0474	0.0191	0.0258
MySQL	0.1080	0.5389	0.359	0.7954

Fig. 6. The results of Test 1

As shown in the figure Fig.6, MongoDB was faster than MySQL in this case, when various operations were executed in parallel. However, the differences between the two databases

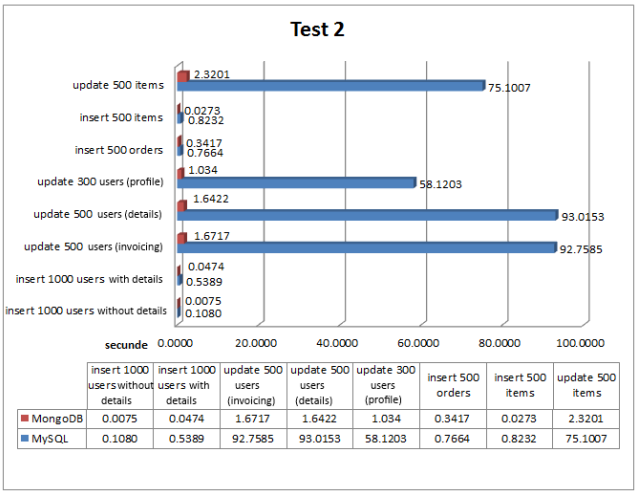
are relatively small, since no operations exceeded one second, in none of the databases.

B. Test 2

The second test includes even more parallel operations, namely:

- insertion of 1.000 users without details (*users* table);
- insertion of 1.000 users with details (*users* table);
- update billing info data for 500 users (*users_billing_info* table);
- update of some users' details, for 500 users (*users_details* table);
- update some users' public profile, for 300 users (*users_public_profile* table);
- insertion of 500 orders (*orders* table);
- insertion of 500 items (*items* orders);
- update of 500 items (*items* table).

The test described above requires simultaneous access to 6 tables, and the results are shown in Fig. 7.



	insert 1000 users without details	insert 1000 users with details	update 500 users (invoicing)	update 500 users (details)	update 300 users (profile)	insert 500 orders	insert 500 items	update 500 items
MongoDB	0.0075	0.0474	1.6717	1.6422	1.034	0.3417	0.0273	2.3201
MySQL	0.1080	0.5389	92.7585	93.0153	58.1203	0.7664	0.8232	75.1007

Fig. 7. The results of Test 2

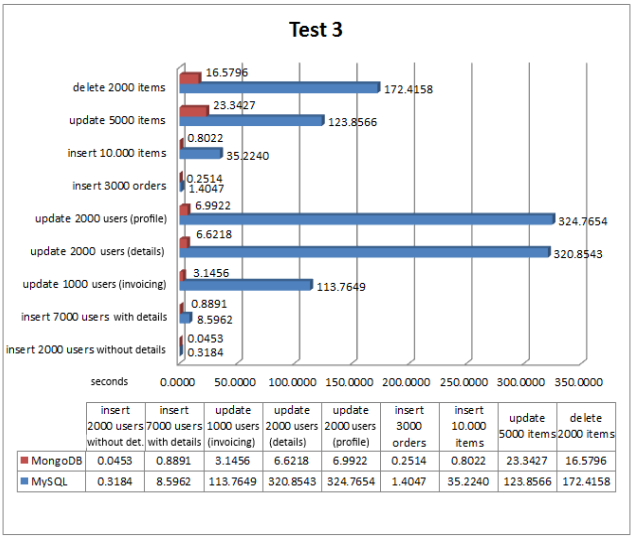
From the results of the second test (Fig 7), we can easily see that the greatest differences between the two databases arise in the *update* operation. While 500 users (with details) have been updated in MySQL in 93 seconds, as many users have been updated in MongoDB in less than 2 seconds. This is due to the different ways of accessing the two databases. For the update operation, the users were randomly selected in both databases.

C. Test 3

Although the operations described in the two tests above highlight the differences between the two databases to conclude we will present further a third test which includes the following:

- insertion of 2.000 users without details;
- insertion of 7.000 users with details;
- update of 5.000 users
  - 1.000 for modifying some billing info data;
  - 2.000 for modifying some general users' details;
  - 2.000 for updating users' public profiles.
- insertion of 3.000 orders;
- insertion of 10.000 de items;
- update of 5.000 items (modify name and description);
- delete of 2.000 items.

The results obtained due to performing this ultimate test are represented in Figure 8 and detailed in the below rows.



	insert 2000 users without det.	insert 7000 users with details	update 1000 users (invoicing)	update 2000 users (details)	update 2000 users (profile)	insert 3000 orders	insert 10.000 items	update 5000 items	delete 2000 items
MongoDB	0.0453	0.8891	3.1456	6.6218	6.9922	0.2514	0.8022	23.3427	16.5796
MySQL	0.3184	8.5962	113.7649	320.8543	324.7654	1.4047	35.2240	123.8566	172.4158

Fig. 8. The results of Test 3

The biggest differences between the two databases have emerged again in the *update* operation, which modified data in both databases. Thus, if 2.000 users change simultaneously different details, MongoDB is faster than MySQL for about 46 times. This is particularly important for applications that use databases that perform constant data update operations, such as Populations Records. For an application that does not require such amount of data to be modified often, such as an online shop, where the insert and delete operations are the most important, the update test is not as relevant as for the application described in this paper.

In a study conducted by authors in 2015 and that was presented in [6], during a conference, the presentation of various case studies in the field of programming and databases, different speed tests in terms of integrating a non-relational database in a forum, were performed. Besides the higher speed that MongoDB had a compared to MySQL, in all the operations that were performed, MongoDB has provided a very important benefit, namely, customizing the application. Thus, due to its use as a database in a forum, the application's structure became very variable, being different for each user. MongoDB allowed modeling the application to the needs of

users, thanks to the fact that a database in MongoDB does not have a predefined data structure, while MySQL has.

IV. CONCLUSIONS

From tests performed and presented for this application, an online publishing platform, the most suitable database was the non-relational MongoDB database. As such, a platform typically has thousands of users or even tens of thousands, MongoDB offered the best solution in terms of the speed at which different operations have to be performed in parallel by many users. If an application that does not require such amount of data to be modified often, such as an online shop, where the insert and delete operations are the most important, the update test is not as relevant as for the application described in this paper.

The advantage of using MongoDB database was further highlighted by conducting the tests and interpreting their results, which were presented in the previous chapter of this paper. MongoDB' query times were much lower than MySQL ones, which is essential when an application has to support thousands of users and multiple operations simultaneously.

We can choose MongoDB instead of MySQL if the application has thousands of users or even tens of thousands, which perform various operations at the same time and the application has to handle a huge volume of data. More and more applications are beginning to use a non-relational database because they provide a more flexible structure; they can support thousands of users and multiple operations simultaneously; they are designed to store large amounts of data and they are denormalized databases, which increases performance.

Switching from a relational database to a non-relational database can be a challenge in many ways, and in the end, the developers have the responsibility to decide which database should be used in a particular application, depending on its requirements and finding the optimal solution for the specific application.

REFERENCES

[1] K. Sanobar, M. Vanita, "SQL Support over MongoDB using Metadata", *International Journal of Scientific and Research Publications*, Volume 3, Issue 10, October 2013

[2] S. Hoberman, "Data Modeling for MongoDB**"**, Publisher by Technics Publications, LLC 2 Lindsley Road Basking Ridge, NJ 07920, USA, ISBN 978-1-935504-70-2, 2014.

[3] H. Martin, "REST and CRUD: the Impedance Mismatch". Developer World. InfoWorld, 29 January 2007, http://www.infoworld.com/article/2640739/application-development/rest-and-crud--the-impedance-mismatch.html.

[4] Kristina Chodorow, "MongoDB: The Definitive Guide, Second Edition" Published by O'Reilly, May 2013, pp 3-4.

[5] T. Frătean, Bazele de date NoSQL – o analiză comparativă, *To Day Software Magazine*, number 10 [Online]. Available: http://www.todaysoftmag.ro/article/304/bazele-de-date-nosql-o-analiza-comparativa

[6] C. Győrödi, R. Győrödi, G. Pecherle, A. Olah, " A comparative study: MongoDB vs. MySQL", 13th International Conference on Engineering of Modern Electric Systems (EMES), 2015 , Oradea, Romania, 11-12 June 2015, pag. 1-6, ISBN 978-1-4799-7650-8.

[7] C. Győrödi, R. Győrödi, R. Sotoc, "A Comparative Study of Relational and Non-Relational Database Models in a Web- Based Application", International Journal of Advanced Computer Science and Applications, Volume 6 Issue 11, 2015, pag. 78-83, ISSN : 2158-107X(Print), ISSN : 2156-5570 (Online).

[8] The definitive guide to PHP's DocBook Rendering System Available: http://www.php.net/, accessed July 2016.

[9] R. P. Padhy, D. Panigrahy, "NoSQL Databases: state-of-the-art and security challenges", International Journal of Recent Engineering Science (IJRES), ISSN: 2349-7157, volume 16, October 2015, http://ijresonline.com/archives/volume-16/IJRES-V16P106.pdf

[10] Yingjie Shi, Xiaofeng Meng, Jing Zhao, Xiangmei Hu, Bingbing Liu, Haiping Wang, "Benchmarking cloud-based data management systems", CloudDB '10 Proceedings of the second international workshop on Cloud data management, pages 47-54, ACM New York, NY,USA 2010, ISBN: 978-1-4503-0380-4, doi: 10.1145/1871929.187193.

[11] N. Jatana, S. Puri, M. Ahuja, I. Kathuria, D. Gosain, "A survey and comparison of relational and non-relational databases", International Journal of Engineering Research &Technology (IJERT) ISSN: 2278-0181, Vol 1, Issue 6, August 2012, pp. 1-5.

Security Risk Assessment of Cloud Computing Services in a Networked Environment

Eli WEINTRAUB
Department of Industrial Engineering and Management
Afeka Tel Aviv Academic College of Engineering
Tel Aviv, Israel

Yuval COHEN
Department of Industrial Engineering and Management
Afeka Tel Aviv Academic College of Engineering
Tel Aviv, Israel

***Abstract*—Different cloud computing service providers offer their customers' services with different risk levels. The customers wish to minimize their risks for a given expenditure or investment. This paper concentrates on consumers' point of view. Cloud computing services are composed of services organized according to a hierarchy of software application services, beneath them platform services which also use infrastructure services. Providers currently offer software services as bundles which include the software, platform and infrastructure services. Providers also offer platform services bundled with infrastructure services. Bundling services prevent customers from splitting their service purchases between a provider of software and a different provider of the underlying platform or infrastructure. In this paper the underlying assumption is the existence of a free competitive market, in which consumers are free to switch their services among providers. The proposed model is aimed at the potential customer who wishes to compare the risks of cloud service bundles offered by providers. The article identifies the major components of risk in each level of cloud computing services. A computational scheme is offered to assess the overall risk on a common scale.**

***Keywords*—Cloud Computing; Risk Management; Information Security; Cloud Risks; Software as a service; Platform as a service; Infrastructure as a service**

I. INTRODUCTION

Traditionally, organizations base their computing facilities on server farms located inside the organization in geographical central sites. In the last years organizations began to shift parts of their computing infrastructures outside the geographic organizational borders to the cloud, where the facilities are owned and managed by other organizations. Reference [1] states that shifting computing infrastructure outside the geographic borders enforces performing changes in production processes and technological changes. Those organizations have to establish new processes of production control, service level monitoring, and resolve security and privacy issues.

Cloud Computing (CC) typically deals with organizations using computing services, communication and web applications. Most definitions state that CC technology enables on-demand services, scalability, and flexibility, in enlarging or downgrading computing consumption ([2] [3]). The National Institute of Standards and Technology (NIST) defines CC as a model for enabling convenient, on-demand network access to a shared pool of configurable computing resources (for example, networks, servers, storage, applications and services) that can be rapidly provisioned and released with minimal management effort or service-provider interaction [4]. Reference [5] argues that occasionally cloud providers suffer outages, thus using a multi-cloud broker is a preferred solution to remove single point of failures. Reference [6] introduces an Inter-Cloud Computing additional layer on top of CC systems which enables shifting resources among the participating cloud systems in occasions of high-load levels.

Cloud computing targets four main groups of organizational customers: private, public, community and hybrid [7]. For private customers, cloud model computing infrastructure services are typically located outside the organization's sites at a cloud service provider. A public customer typically chooses cloud service providers through a bidding process, issuing request for proposal, choosing the best proposal, and contracting for the best bidder having the best proposal. The cloud computing provider may use the same computing infrastructure for supplying the needs of other companies. In a community model, infrastructure services are shared by a group of customers. In a hybrid model, an organization can use infrastructure services supplied by public, private or as part of a community. Reference [8] researched the emerging themes in financial services technologies and found that cloud computing seems to be a cost-effective infrastructure affording capital efficiency for financial services providers.

This article reviews the main motivations and obstacles to adopting the cloud technology by companies. Information security has been found as a barrier to CC adoption, and is an issue dealt intensively in CC research [9]. Reference [10] researched CC trends, claims that security will not be a barrier for cloud adoption, since it will be implemented by centralized automated processes.

This article is organized as follows: Section II is an overview of the current CC architecture and the dynamic networked architecture which is used by this paper. Section III is an overview of risk management theory. Section IV overviews security risks prevalent in CC architecture. Section V presents the risk optimization proposed model. Section VI discusses the possible CC architectures for implementation of the model including formulation and a case study illustrating the model. Finally section VII concludes the advantages of the model and future possible research.

II. Cloud Computing Architecture

Cloud computing architecture is described in literature as consisting of three layers: IaaS, PaaS and SaaS. Each layer performs certain functions, serving consumers' requests and also supporting functions requested by upper layers. This separation to layers also fits current services offered by CC providers. Reference [7] defines a framework of CC architecture composing three layers of functions supporting cloud computing services. Fig. I describes architectures' components. Rectangles describe computing services. The business buys all cloud services from one SP.

Following the functions performed by each layer.

Infrastructure layer – This layer focuses on providing technologies as basic hardware components for software services. There are two kinds of infrastructures: storage capabilities and computing power.

Platform layer - includes services which are using cloud infrastructures needed for their functioning. There are two kinds of platform services: development and business platforms. Development platforms are aimed for usage by developers who write programs before transferring them to production and usage by organizations' users. Business platforms enable organizational developers make adaptations of software packages for deployment in their organizations.

Application layer - consists of the programs and human interfaces used by the organizations' end-users. Applications are running on cloud assets, making use of platform and infrastructure layers. There are two kinds of services in this layer: applications and on-demand services. Application services are software packages ready for end-users such as Microsoft Office, while on-demand services are software applications which are used by the organizations' customers. Those services are used according to on-demand needs, and used on a pay-per-use or fixed-price pricing model.

Service Providers (SP) offer their customers three kinds of services: IaaS, PaaS and SaaS. Each SP manages all underlying infrastructure for the offered service. For example a SP suggesting a SaaS product is also bundling into the product the PaaS and IaaS layers. Reference [11] states that according to cloud computing architecture a certain provider may run an application using another provider's infrastructure, but in practice both providers are parts of the same organization. Current practice is that when a provider suggests selling a PaaS service he also bundles the IaaS layer in the deal. Such bundling by service providers limit free market forces from entering the competition, forcing customers pay for components they may buy cheaper from other providers. For example a customer may buy a SaaS service from SP1, but buy the underlying PaaS service from SP2 which sells the appropriate platform service cheaper than SP1. Reference [12] claims that in the future, developers will plan their cloud applications which will enable migration of services among clouds of multiple clouds. According to [11] cloud computing architecture is more modular compared to traditional hosting architectures based in server farms, and programs running on different layers are loosely coupled, thus enabling the development of a wide range of applications. Reference [2] also claims that it is possible that applications belonging to different layers will be run on separate geographical locations even in different countries. Reference [13] claims that virtual machine migration allows transfer of a running application from one virtual machine to another, which may be provided by a different IaaS provider. Reference [14] proposes to make use of multiple distinct clouds simultaneously thus achieving security merits by making use of multiple distinct clouds simultaneously. This article continues the research direction proposed in [15] basing CC services on a dynamic business model which enables implementing functionalities of a service provider interfacing the underlying platform or infrastructure service by other service providers according to consumers' preferences. References [16] [15] demonstrate added values achieved in aspects of consumers' cost optimization and consumers' utility optimization. This research is aimed at suggesting a new technique for risk assessment which minimizes risks, utilizing the dynamic CC architecture. Implementing this required functionality puts two requirements on cloud architecture. The architecture should be based on open standards which will enable interfacing between many components among all providers in all three layers. Second, the architectures' building blocks should be loosely coupled. Implementation of those two functionalities should enable connectivity among vertical and horizontal services, thus eliminating the bundling phenomena. Figure II describes the dynamic CC architecture. Arrows describe services supplied by underlying layers. Rectangles describe cloud computing services. The business consumes its CC services from many SP's choosing the best combination of service providers.

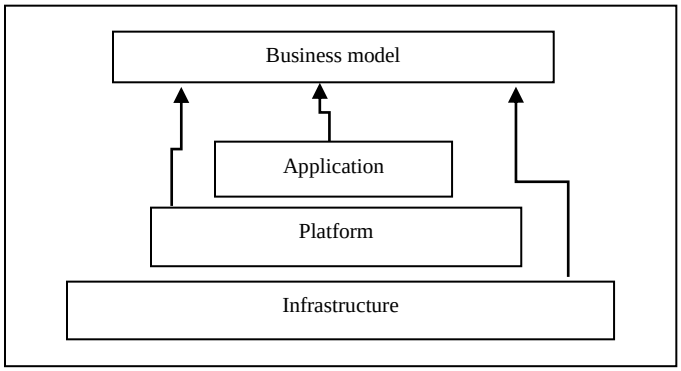


Fig. 1. Current Cloud business model Architecture – One SP

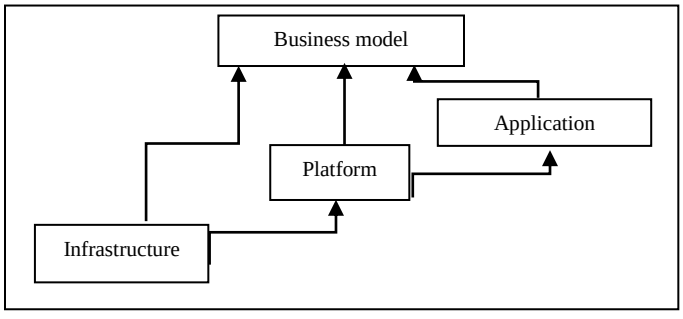


Fig. 2. A Dynamic Architecture for Cloud Computing – Many SP's

III. Cloud Computing Security Risks

Researchers state that security challenges are among the biggest obstacles to adoption cloud services [14]. Reference [17] states that CC as the most prevalent IT outsourcing paradigm still entails serious IT security risks, and also states that researchers still are not able to fully capture the complex nature of IT security risks and how to measure it. Industry research and advisory company IDC analysts report that 87.5% of their members indicate that cloud security is their number 1 challenge [18]. Reference [19] States that managing security risks to business systems is getting more and more complex and time consuming, and many publications include proposals targeting the various cloud security threats. Following, an overview of research published in the cloud computing security risks domain.

Cloud security covers several categories. Reference [20] surveyed the research publications on cloud security issues, addressing vulnerabilities, threats, and attacks. In order to understand security risks, the authors identify the basic concepts underlying vulnerabilities and threats, and classify them as follows: virtualization elements, multi-tenancy, cloud platform and software, data outsourcing, data storage security and standardization and trust. The authors then address the security risks and topics involved in managing risks of each category. Reference [21] states that cloud threats are due to the complex virtualized infrastructure and dynamic nature of the cloud and they can be categorized to three kinds: (1) Multiple Users – A virtualized cloud layer such as IaaS can hold up various virtual machines and can provide multiple access to different users from around the globe, this kind of sharing is responsible for information leakage. (2) Minimal Control – Users of the cloud are not aware of the location of the physical server, as all these physical servers belong to the data centers of the providers hence the users are not aware of the location of their VMs and the provider is not aware of the contents of the VM or its applications hence giving a way to the security threats. (3) Single Point of control - All the virtualized servers are connected to one or limited number of network interface cards (NIC). This in turn causes more vulnerabilities in the virtual environment, any compromise to the security of the VMs or the physical server will lead to the compromise of either the VM or the physical server and will enable the hacker to gain access to either physical server. Reference [22] presents the results of a case study identifying real-world information security documentation issues for a Global Fortune 500 organization, should the organization decide to implement cloud computing services in the future. According to [22] security risks can be categorized to the following domains: Governance and Enterprise Risk Management; Legal Issues; Compliance and Audit Management; Information Management and Data Security; Interoperability and Portability; Traditional Security, Business Continuity and Disaster Recovery; Data Centre Operations; Incident Response; Application Security; Encryption and Key Management; Identity, Entitlement and Access Management; Virtualization. CSA's experts identified nine critical threats, ranked in descending order of severity: Data Breaches, Data Loss, Account Hijacking, Insecure APIs, Denial of Service, Malicious Insiders, Abuse of Cloud Services, Insufficient Due Diligence, and Shared Technology Issues. This list of threats could serve as a guide to help users and providers make decisions about risk mitigation in their organizations [23]. Reference [17] proposes a comprehensive conceptualization of Perceived IT Security Risks in the CC context that is based on six distinct risk dimensions grounded on an extensive literature review, Q-sorting, and expert interviews. Second, a multiple-indicators and multiple-causes analysis of data collected from 356 organizations is found to support the proposed conceptualization as a second-order aggregate construct. The final set of six security risk dimensions is: Confidentiality, Integrity, Availability, Performance, Accountability and Maintainability risks. Each risk dimension is further categorized to risk items, in total 31 risk items. For example performance risk is categorized to network risks, scalability risks, underperformance risks and internal performance risks. Reference [19] presents a method to assess security risks including a cohesive set of steps to identify a complete set of security risks and also to assess them. The method is based on the integration of qualitative and quantitative models that focus on formal evaluation and assessment. In order to assess risks, risks are categorized to Six-View Perspectives: Threat view, Resource View, Process View, Risk Assessment View, Management View, and Legal View. To summarize, there is no one single framework describing all CC risk factors.

This paper follows ISACA's framework defined in [24]. The framework is designed to present practical guidance and facilitate the decision process for IT and business professionals concerning the decision to move to the cloud. The guide provides checklists outlining the security factors to be considered when evaluating the cloud as a potential solution. Evaluating cloud-related risks raises the need to define the information assets needing protection. Assets can be categorized to data, applications and processes. The impact of a migration to the cloud depends on the cloud service model and deployment model being considered. The combination of service model and deployment model can help identify an appropriate balance for organizational assets.

These assets are commonly subject to the following risk events:

- **Unavailability**—The asset is unavailable and cannot be used or accessed by the enterprise.
- **Loss**—The asset is lost or destroyed.
- **Theft**—The asset has been intentionally stolen and is now in possession of another individual/enterprise. Theft is a deliberate action that can involve data loss.
- **Disclosure**—The asset has been released to unauthorized staff/enterprises/organizations or to the public. This also includes the undesired, but legal, access to data due to different regulations across international borders.

IV. Risk Modeling

Starting our analysis we note that the damage of loss is greater than the damage of unavailability. Also, disclosure mainly pertains to data. Finally, the risk of theft means unavailability, and includes the risks of both loss and

disclosure. It is therefore important to map the implications of these relationships as shown in figures III for applications and processes, and figure IV for data.

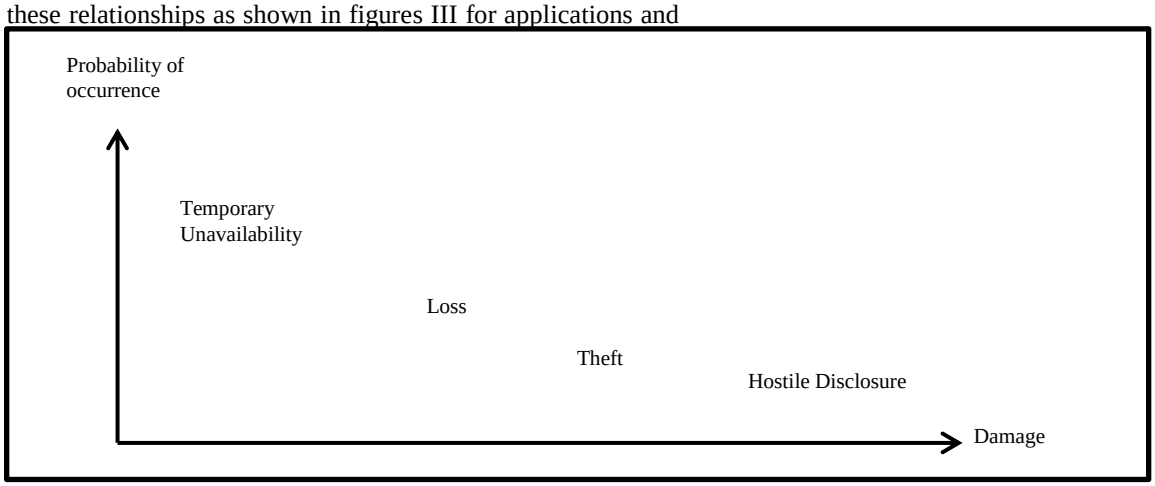


Fig. 3. Qualitative characterization of applications and process risks

In terms of policy, in some cases the damage of temporary unavailability (of process or application) is so minor as to ignore it altogether. In case of process or application the theft damage is usually too small to justify insurance.

As to data: it its risk is mainly depends on the data's criticality as shown in figure IV.

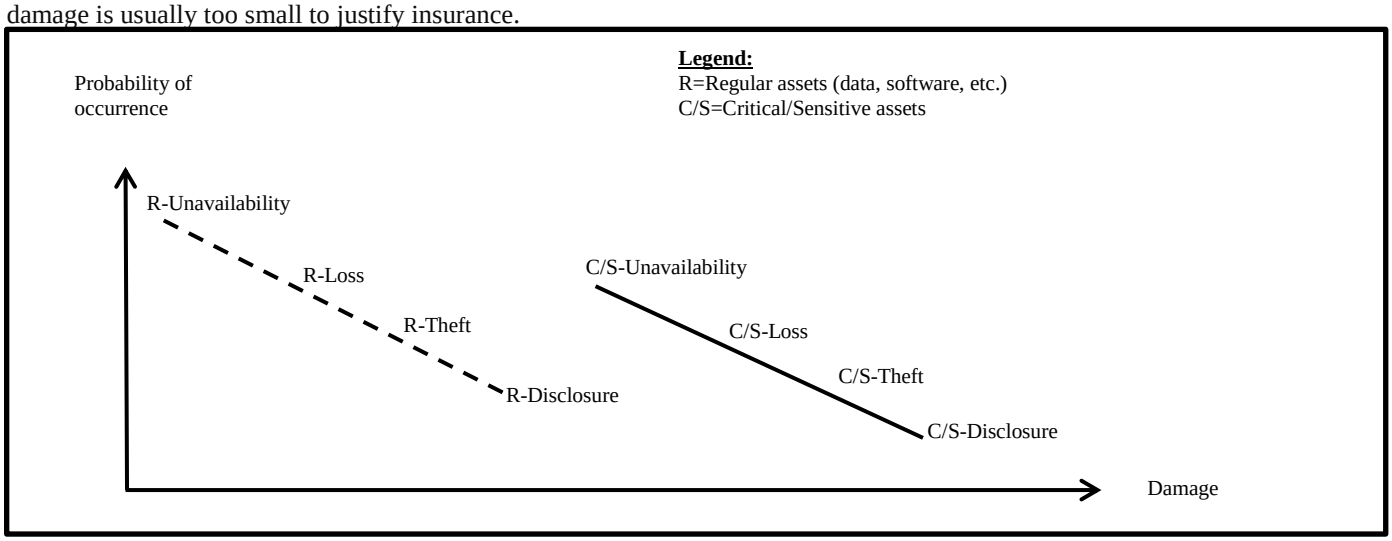


Fig. 4. Typical qualitative characterization of regular and critical organizational data risks

Naturally, the damage due to critical data risks is larger than the damage to regular data risks. The critical data are typically more protected and therefore its failure probabilities in figure IV are lower than those of the regular data. Note that disclosure of regular data has low risk, while the disclosure of critical data has higher risk. Also, temporary unavailability of regular data may be tolerated due to low damage. On the other hand, theft of critical data is typically insured.

The following discussion will lead to the assessment of risks that stem from possible damages and the occurrence probabilities/rates. The result is a set of weights according to which comparison of risks could be made.

Next, [24] published a list of CC risk factors. Common risk factors that are not linked solely to cloud infrastructures, but apply to all types of infrastructure, are not covered in the list. Examples of such risk factors include external hacking, malicious insiders, mobile computing vulnerabilities, virus and malicious code and business impact due to provider inability. Following the list of risks, categorized to the three cloud layers, each risk includes an indication for either risk increasing (RI) or risk decreasing (RD). Additionally the description includes the types of risks and their severity level. From figure IV it is clear that the risk level increases from Unavailability to Loss, from Loss to Theft, and from Theft to Disclosure. While assigning values to severity levels may be an open issue, the following example is an arbitrary scheme where these values are increasing as mentioned above.

Example: IaaS Risk Grade Computations and SPs comparison

The proposed model intentionally takes a generalization approach to bypasses the details of the effects of each combination of: (1) factor, (2) risk-type, and (3) supplier.

Instead, for each of the three layers (IaaS/PaaS/SaaS) two aggregated risk grades are computed: (1) Risk Increasing (RI) and (2) Risk Decreasing (RD). RI is related to a factor group that is partly oriented to exposure factors, whereas RD is more related to protection factors.

Table I not only differentiate between the RI/RD factors, it also depicts the probability and damage of each factor to the four risk types (Unavailability, Loss, Theft, and Disclosure). The risk level of each risk type is a measure of both the probability of occurrence and the expected damage of the risk realization. In the example, the probability and the damage are ranked on a 5 point scale (1 to 5) and the risk is the multiplication of the probability rank and the damage rank. For each factor, each risk type is evaluated through such a multiplication yielding a scale of 1 through 25. In this way the risk level of each risk factor is computed by summing the relevant risk values over the 4 risk types (Unavailability, Loss, Theft, and Disclosure). For example, in Table I: the first factor "Legal transcoder requirements" affects only the disclosure risk type having a probability rank of 2 and a damage of 4 which reflects risk level of 8. On the other hand the second risk factor "Multitenancy and isolation failure" is relevant to both theft (risk=3) and disclosure (risk=4) and therefore its risk level is 3+4=7.

TABLE I. EXAMPLE OF COMPUTING IMPORTANCE WEIGHTS FOR IAAS RISK FACTORS (VALUES ARE ARBITRARY-FOR ILLUSTRATION ONLY)

P=Probability D=Damage R=Risk level	Unavailability			Loss			Theft			Disclosure		
IAAS: (RI) Risk increasing factors	*P* 1-5	*D* 1	*R=P*D* Risk	*P* 1-5	*D* 2	*R=P*D* Risk	*P* 1-5	*D* 3	*R=P*D* Risk	*P* 1-5	*D* 4	*R=P*D* Risk
A. Legal transborder requirements										2	4	2*4=8
B. Multitenancy and isolation failure							1	3	1*3=3	1	4	1*4=4
C. Lack of visibility of technical security measures	3	1	3*1=3	3	2	3*2=6	3	3	3*3=9	3	4	3*4=12
D. Absence of DRP and backup	3	1	3*1=3	4	2	4*2=8						
E. Physical security							3	3	3*3=9	3	4	3*4=12
F. Data disposal										2	4	2*4=8
G. Offshoring infrastructure	2	1	2*1=2	2	2	2*2=4	2	3	2*3=6	4	4	3*4=16
H. Virtual machine (VM) security maintenance	3	1	3*1=3	2	2	2*2=4	2	3	2*3=6	3	4	3*4=12
I. Cloud provider authenticity	2	1	2*1=2	3	2	3*2=6	3	3	3*3=9	3	4	3*4=12
IAAS : (RD) Risk Decreasing factors	*P* 1-5	*D* 1	*R=P*D* Risk	*P* 1-5	*D* 2	*R=P*D* Risk	*P* 1-5	*D* 3	*R=P*D* Risk	*P* 1-5	*D* 4	*R=P*D* Risk
J. Scalability and elasticity	2	1	2*1=2					3				
K. DRP and backup	4	1	4*1=4	4	2	4*2=8	3	3	3*3=9			
L. Patch management	2	1	2*1=2	3	2	3*2=6	4	3	4*3=12	2	4	2*4=8

The risks of Table I are summarized in table II. Table II aggregates the risks of each risk factor and gives them weights proportional to their contribution to the total risk of the category (Category is defined by one of I/P/S and one of RI/RD). Once all the risk levels of a category (RI, or RD) are known they are summarized and each risk factor % weight is computed as the portion it contributes to the total risk level. For example, the IaaS RI risk category sums to 167 so the weight of first risk factor "Legal transcoder requirements" is computed as the ratio of its risk level (8) to the total RI risk (167) – so % RI weight=8/167=5%.

TABLE III. SUMMARY OF IAAS RISK FACTORS FROM TABLE I.

IAAS: (RI) Risk increasing factors	*Availability* Risk	*Loss* Risk	*Theft* Risk	*Disclosure* Risk	*Total risk*	*% of total*
A. Legal transborder requirements				2*4=8	8	5%
B. Multitenancy and isolation failure			1*3=3	1*4=4	7	4%
C. Lack of visibility of technical security measures	3*1=3	3*2=6	3*3=9	3*4=12	30	18%
D. Absence of DRP and backup	3*1=3	4*2=8			11	7%
E. Physical security			3*3=9	3*4=12	21	12%
F. Data disposal				2*4=8	8	5%
G. Offshoring infrastructure	2*1=2	2*2=4	2*3=6	3*4=16	28	17%
H. Virtual machine (VM) security maintenance	3*1=3	2*2=4	2*3=6	3*4=12	25	15%
I. Cloud provider authenticity	2*1=2	3*2=6	3*3=9	3*4=12	29	17%
Total	**13**	**28**	**42**	**84**	**167**	**100%**
IAAS : (RD) **Risk Decreasing factors**	*R=P*D* Risk	*R=P*D* Risk	*R=P*D* Risk	*R=P*D* Risk		
J. Scalability and elasticity	2*1=2				2	4%
K. DRP and backup	4*1=4	4*2=8	3*3=9		21	41%
L. Patch management	2*1=2	3*2=6	4*3=12	2*4=8	28	55%
Total	**8**	**14**	**21**	**8**	**51**	**100%**

Thus, the IaaS importance weights appear on the right hand side of table II.

The next step is to grade the factor list of the alternative SPs. The grades are based on a 0 to 100 quality scale for each factor: where the best (minimum risk) =100 and the worst (maximum risk) =0. This is kept consistent in grading of both RI factors and RD factors. Therefore, it is desirable to get high grades in both RI and RD factors.

Table III describe an example of grading of 3 theoretical Service Providers (SPs).

TABLE IV. EXAMPLE OF IAAS RISK COMPARISON OF 3 CC SPS

IAAS: (RI) **Risk increasing factors**	**Grades SP - 1**	**Grades SP - 2**	**Grades SP -3**	**% RI Importance**	**Grade-1**	**Grade-2**	**Grade-3**
A. Legal transborder requirements	90	93	88	**5%**	4.5	4.7	4.4
B. Multitenancy and isolation failure	67	75	83	**4%**	2.7	3.0	3.3
C. Lack of visibility surrounding technical security measures in place	98	91	64	**18%**	17.6	16.4	11.5
D. Absence of DRP and backup	82	88	95	**7%**	5.7	6.2	6.7
E. Physical security	76	87	90	**12%**	9.1	10.4	10.8
F. Data disposal	69	74	72	**5%**	3.5	3.7	3.6
G. Offshoring infrastructure	95	87	88	**17%**	16.2	14.8	15.0
H. Virtual machine (VM) security maintenance	81	70	79	**15%**	12.2	10.5	11.9
I. Cloud provider authenticity	100	96	77	**17%**	17.0	16.3	13.1
Total				**100%**	88	86	80
IAAS : (RD) **Risk Decreasing factors**	**Alternative -**	**Alternative - 2**	**Alternative -3**	**% RD Importance**	**Grade-1**	**Grade-2**	**Grade-3**
J. Scalability and elasticity	63	98	100	**4%**	2.5	3.9	4.0
K. DRP and backup	69	91	93	**41%**	28.3	37.3	38.1
L. Patch management	84	90	84	**55%**	46.2	49.5	46.2
Total				**100%**	77	91	88

Thus, each SP has two graded components for the IaaS risk: (RI, RD)

SP1: (88, 77);

SP2: (86, 91);

SP3: (80, 88);

In cases where selecting a SP to a certain layer would be independent of the selection of SP to other layers we could decide on the SP on the basis of the above grading. For example, a comparison shows that SP2 dominates SP3 (86>80, 91>88), and has a trade-off of (RI= -2, RD=14). Therefore, SP2 is the better choice as long as: (importance (RI)/importance (RD)) ≤7.

Continued example: PaaS and SaaS SP comparisons

The same procedure illustrated on IaaS is performed on the factors of PaaS and SaaS. To continue the example Tables IV and V show only the last part of comparing the three different

alternative SPs in each of the layers. Platform risks are weighted separately

TABLE V. EXAMPLE OF PAAS RISK COMPARISON OF 3 CC SPS

PAAS: (RI) Risk increasing factors	***Grades SP - 2***	***Grades SP - 2***	***Grades SP -3***	**% RI Importance**	**Grade-1**	**Grade-2**	**Grade-3**
A. Application mapping	63	63	82	29%	18.3	18.3	23.8
B. SOA-related vulnerabilities	61	82	84	42%	25.6	34.4	35.3
C. Application disposal	74	76	97	29%	21.5	22.0	28.1
Total				**100%**	65	75	87
PAAS : (RD) Risk Decreasing factors	***Alternative-1***	***Alternative - 2***	***Alternative -3***	**% RD Importance**	**Grade-1**	**Grade-2**	**Grade-3**
D. Short development time	94	93	86	46%	43.2	42.8	39.6
E. Platform security features	95	80	94	54%	51.3	43.2	50.8
Total				100%	95	86	90

Thus each SP has two graded components for the PaaS risk: (RI (Increasing), RD (Decreasing)):

SP1: (65, 95);

SP2: (75, 86);

SP3: (87, 90);

Considering PaaS grades, **SP3** dominates **SP2** (87>75 and 90>86), so SP2 is not a relevant candidate. The comparing SP3 to SP1 gives a trade-off: (18,-5) which favors **SP3** as long as (importance (RD)/(importance(RI)≤3.4).

TABLE VII. EXAMPLE OF SAAS RISK COMPARISON OF 3 CC SPS

SAAS: (RI) Risk increasing factors	**Grades SP-1**	**Grades SP - 2**	**Grades SP -3**	**% RI Importance**	**Grade-1**	**Grade-2**	**Grade-3**
A. Data ownership	88	70	97	**11**	9.7	7.7	10.7
B. Data disposal	92	65	79	**11**	10.1	7.2	8.7
C. Lack of visibility into software systems development life cycle (SDLC)	92	66	88	**16**	14.7	10.6	14.1
D. Identity and access management (IAM)	97	83	85	**15%**	14.6	12.5	12.8
E. Exit strategy	89	66	90	**5%**	4.5	3.3	4.5
F. Broad exposure of applications	92	80	73	**11%**	10.1	8.8	8.0
G. Ease to contract SaaS	76	95	67	**15%**	11.4	14.3	10.1
H. Lack of control of the release management process	75	67	66	**5%**	3.8	3.4	3.3
I. Browser vulnerabilities	94	68	67	**11%**	10.3	17.6	13.4
Total				**100%**	89	85	85
SAAS : (RD) Risk Decreasing factors	**Grades SP-1**	**Grades SP - 2**	**Grades SP -3**	**% RD Importance**	**Grade-1**	**Grade-2**	**Grade-3**
D. A Improved security	62	77	86	**50%**	31.0	38.5	43.0
E. Application patch management	99	84	91	**50%**	49.5	42.0	45.5
Total				**100%**	81	81	89

Thus each SP has two graded components for the SaaS risk: (RI, RD):

SP1: (89, 81);

SP2: (85, 81);

SP3: (85, 89);

Considering SaaS grades, **SP2** is dominated by both **SP3** (85=85 but 89>81), and by SP1 (89>85 and 81=81) so **SP2** is not a relevant candidate. However, the trade-off between SP3 and SP1 is: (-4, 8) so SP3 would be preferred as long as the importance of RD is more than half the importance of RI. Else, SP1 would be selected.

Risk assessment in two Cloud Computing Architectures: One SP versus Many SP's.

In this section we compare two different scenarios: The first scenario is where SP's bundle their offerings in the three layers, consequently a choice of a single SP must be made. This scenario is implemented on the Current Cloud business model Architecture – One SP described in Fig. I. This scenario will lead to choose the least risky SP. The second scenario is

where the competition and free market forces are leading so that services could be purchased independently for each of the three CC layers (infrastructure, platform and software). This scenario is implemented on the Dynamic Architecture for Cloud Computing – Many SP's, described in Fig. II.

For the case where SPs bundle their services (as in current practices) the assumption would be different. In such a case, each SP has the full chain of three layers to offer. Since the risk of any chain is reflected by the chain's most vulnerable point, it is conceivable to grade the SPs by their minimum risk levels.

For example, SP1 risk grades are: IaaS (88, 77), PaaS (65, 95) SaaS (89, 81).

Therefore the grades for SP1 are: RI = Min{88, 65, 89) = 65; RD = Min{77, 95, 81) = **77**, yielding SP1 grade = (65,77). Computations for the bundling case in table IV. So the overall SP grades would be:

TABLE VIII. RISK COMPUTATIONS FOR THE BUNDLING EXAMPLE

IaaS	PaaS	SaaS	Overall SP grade
SP1: (88, 77);	SP1: (65, 95);	SP1: (89, 81);	SP1 grade = (65, 77)
SP2: (86, 91);	SP2: (75, 86);	SP2: (85, 81);	SP2 grade = (75, 81)
SP3: (80, 88);	SP3: (87, 90);	SP3: (85, 89);	SP3 grade = (80, 88)

Thus, under the assumptions of layers bundling and a single SP selection it is clear that SP3 dominates S2 which dominates SP1.

This is true for both the RI grades: 80>75>65, and the RD grades: 88>81>77.

So SP3 is selected with SP3 grade = **(80, 88).**

Under convergence to the free market competition, each layer would be independently selected. In this case customers choose the best SP for each layer independent of their decisions in other layers. Computations for the free market example in table VII.

TABLE IX. RISK COMPUTATIONS FOR THE FREE MARKET EXAMPLE

	IaaS	PaaS	SaaS
	SP1: (88, 77);	SP1: (65, 95);	SP1: (89, 81);
	SP2: (86, 91);	SP2: (75, 86);	SP2: (85, 81);
	SP3: (80, 88);	SP3: (87, 90);	SP3: (85, 89);
Selected SP(grade):	**SP2 (86, 91);**	**SP3 (87, 90);**	**SP3 (85, 89);**

Thus, the maximal risk management solution leads to choosing SP2 for IaaS, and SP3 for PaaS and SaaS.

For this selection RI = Min{86,87,85} = 85; RD = Min{91,90,89} = 89

Thus, the overall grade is: **(85, 89)** which is better and dominates the single SP3 grade = **(80, 88)**

To conclude the first example, it has been demonstrated that the Dynamic proposed CC architecture enables achieving higher risk scores then the traditional one-SP model by choosing a combination of services offered by several CC SP's.

Second Example

It should be clear that once the importance percentages of various service items were set (as in tables I, II) they will stay constant for quite a while, and change only when overall revision is needed. On the other hand, the grades for these service items may change in time for certain suppliers and some new suppliers may join the competition.

Let us assume that two years after the grades above were computed a new decision point comes along and the new grades along with those of two new suppliers are now summarized in table VIII.

TABLE X. RISK COMPUTATIONS FOR THE SECOND BUNDLING EXAMPLE

IaaS	PaaS	SaaS	Overall SP grade
SP1: (81, 75);	SP1: (70, 90);	SP1: (92, 85);	SP1 grade = (70, 75)
SP2: (87, 92);	SP2: (75, 85);	SP2: (80, 85);	SP2 grade = (75, 85)
SP3: (85, 85);	SP3: (87, 90);	SP3: (75, 90);	SP3 grade = (75, 85)
SP4: (85, 80);	SP4: (80, 80);	SP4: (90, 90);	SP4 grade = (80, 80)
SP5: (80, 90);	SP5: (85, 85);	SP5: (85, 90);	SP5 grade = (80, 85)

Thus, under the assumptions of layers bundling and a single SP selection it is clear that SP5 dominates all the other SPs.

This is true for both the RI grades: 85>80>75>70, and the RD grades: 85>80>75.

So SP5 is selected at this point in time with SP5 grade = **(80, 85).**

Under convergence to the free market competition, each layer would be independently selected. In this case customers choose the best SP for each layer independent of their decisions in other layers. Computations for the free market example in table IX.

TABLE XI. RISK COMPUTATIONS FOR THE SECOND FREE MARKET EXAMPLE

	IaaS	PaaS	SaaS
	SP1: (81, 75);	SP1: (70, 90);	SP1: (92, 85);
	SP2: (87, 92);	SP2: (75, 85);	SP2: (80, 85);
	SP3: (85, 85);	SP3: (87, 90);	SP3: (75, 90);
	SP4: (85, 80);	SP4: (80, 80);	SP4: (90, 90);
	SP5: (80, 90);	SP5: (85, 85);	SP5: (85, 90);
Selected SP(grade):	**SP2 (87, 92)**	**SP3 (87, 90)**	**SP4 (90, 90)**

Thus, the maximal risk management solution leads to choosing SP2 for IaaS, SP3 for PaaS, and SP4 for SaaS.

For this selection RI = Min{87,87,90} = 87; RD = Min{92,90,90} = 90

Thus, the overall grade is: **(87, 90)** which is better and dominates the single SP5 grade = **(80, 85)**.

This example shows that organizations should follow the decision process finding the best solution each time new SP enters the market, improving their risk grades. In addition, the dynamic model enables achieving improved risk grades over the traditional One-SP model.

V. CONCLUSIONS

This paper proposes a technique for evaluating and comparing risks between different service providers in the three CC layers. The technique is illustrated through a numeric example which also shows the advantage of free market competition, where purchasing services independently for each layer leads to a superior choice with least risk exposure.

Two preconditions are required for effective competition, and for our risk assessment models to be effective. We claim market forces are bound to cause these conditions to materialize in the long run. First, suppliers have to offer standard features of their services since comparing risk probabilities/damages has to relate to similar functionalities. This will be the ground for a comparison of dimensional risk scores relating to similar services. Second, software suppliers should build their services according to open standards, (which nowadays are not the case), thus enabling connectivity among different services offered by suppliers.

Future research directions may span the following directions: 1. Calculating risk according to specific proportional weights assigned to risk increasing versus risk decreasing factors up to consumers' risk appetite. 2. Enhancing the proposed technique to compute the optimized solutions by finding the risk increasing/decreasing proportion which bring the minimal risk. 3. Add deployment risk factors to risk computations as suggested by [24].

The proposed risk assessment model could be elaborated to incorporate the connectivity costs among different SP's. Interfacing a specific service between two SP's needs budget investments in the first establishing of the interface and in the ongoing budgetary expenses depending on service consumption. This raises the need for a multi-objective risk assessment model which takes into consideration optimizing risk assessment under budget costs.

REFERENCES

[1] T. Pueschel, A. Anandasivam, S. Buschek, and D. Neumann, "Making money with clouds: Revenue optimization through automated policy decisions". ECIS - European Conference on Information Systems 17, 2009.

[2] A. Velte, R. Elsenpeter, and T. J. Velte, "Cloud Computing: A practical approach". Tata McGraw-Hill Education Pvt. Ltd, 2009.

[3] L. Vaquero, L. Rodero-Merino, J. Caceres, and M. Lindner, "A Break in the Clouds: Towards a Cloud Definition. Editorial note". ACM SIGCOMM (2009). Computer Communication Review 50 Volume 39, Number 1, January 2009.

[4] P. Mell, and T. Grance, "The NIST definition of cloud computing", National Institute of Standards and Technology, NIST, Vol. 53 No. 6, p. 50, 2009.

[5] Y. Mansouri, A. N. Toosi, and R. Buyya, "Brokering Algorithms for Optimizing the Availability and Cost of Cloud Storage Services", IEEE International Conference on Cloud Computing Technology and Science, 2013.

[6] T. Aoyama, and H. Sakai, "Inter-Cloud Computing", Business Information Systems Engineering, March 2013.

[7] C. Weinhardt, B. Blau, and J. Stößer, "Cloud Computing – A Classification, Business Models, and Research Directions". Business & Information Systems Engineering, May 2009.

[8] A. Gill, D. Banker, and P. Seltsika, "Moving Forward: Emerging Themes in Financial Services Technologies Adoption", Communications of the Association for Information Systems: Vol. 36, Article 12, 2015.

[9] Z. Chen, F. Han, J. Cao, X. Jiang, and S. Chen, "Cloud Computing-Based Forensic Analysis for Collaborative Network Security Management System", Tsinghua science and technology, Vol 18/1, February 2013.

[10] J. Staten, "Forrester, Cloud predictions for 2014: Cloud joins the IT portfolio", http://blogs.forrester.com/james_staten/13-12-04-cloud_computing_predictions_for_2014_cloud_joins_the_formal_it_portfolio, accessed 02 March 2014.

[11] Q. Zhang, L. Cheng, and R. Bautaba, "Cloud computing: State-of-the-art and Research challenges", J Internet Serv Appl 1:7-18, 2010.

[12] F. Paraiso, N. Haderer, P. Merle, R. Rouvoy, and L. Seinturier, "A Federated Multi-Cloud PaaS Infrastructure", IEEE Fifth International Conference on Cloud Computing, 2012.

[13] U. Z. Rehman, F. K. Hussain, and O. K. Hussain, "Towards Multi-Criteria Cloud Service Selection", Fifth International Conference on Innovative Mobile and Internet Services in Ubiquitous Computing, 2011.

[14] J. Bohli, N. Gruschka, M. Jensen, L.L. Iacono, and N. Marnau, "Security and Privacy-Enhancing Multi cloud Architectures", IEEE Transactions on Dependable and Secure Computing, Vol. 10, No' 4, 2013.

[15] E. Weintraub and Y. Cohen, "Cost Optimization of Cloud Computing Services in a Networked Environment", (IJACSA) International Journal of Advanced Computer Science and Applications ,Vol. 6, No. 4, pp. 148-157, 2015.

[16] E. Weintraub and Y. Cohen, "Optimizing User's Utility from Cloud Computing Services in a Networked Environment", (IJACSA) International Journal of Advanced Computer Science and Applications, Vol. 6, No. 10, pp. 153-163, 2015.

[17] T. Ackermann, T. Widjaja, A. Benlian, and P. Buzmann, "Percieved IT Security Risks of Cloud Computing: Conceptualization and Scale

Development, Thirty Third International Conference on Information Systems, Orlando USA, 2012.

[18] C. A. Christiansen, C. J. Kolodgy, S. Hudson, and G. Pintal, IDC – White paper – "Identity and Access Management for Approaching Clouds", May 2010.

[19] S. B. Yadav, and D. Tianxi, "A Comprehensive Method to Assess Work System Security Risk," Communications of the Association for Information Systems: Vol. 34, Article 8, 2014.

[20] D. A. B. Fernandes, L. F. B. Soares, J. V. Gomes, M. M Freire and P. R. M. Inácio, "Security issues in cloud environments: a survey", Int. J. Inf. Secur. 13:113–170, 2014.

[21] B. Mansukhani and T. A. Zia, "The Security Challemges and Countermeasures of Virtual Cloud", Australian Information Security Management Confference, 2012.

[22] G. Grispos, W. Glisson, and T. Storer, "Cloud Security Challenges: Investigating Policies, Standards, and Guidelines in A Fortune 500 Organization", ECIS, 2013.

[23] CSA - Cloud Security Alliance, "The Notorious Nine Cloud Computing Top Threats in 2013", USA, 2013.

[24] ISACA, "Security Considerations for Cloud Computing", USA, 2012.

Using Multiple Seasonal Holt-Winters Exponential Smoothing to Predict Cloud Resource Provisioning

Ashraf A. Shahin[1,2]
[1]College of Computer and Information Sciences,
Al Imam Mohammad Ibn Saud Islamic University (IMSIU)
Riyadh, Kingdom of Saudi Arabia
[2]Department of Computer and Information Sciences, Institute of Statistical Studies & Research,
Cairo University,
Cairo, Egypt

***Abstract*—Elasticity is one of the key features of cloud computing that attracts many SaaS providers to minimize their services' cost. Cost is minimized by automatically provision and release computational resources depend on actual computational needs. However, delay of starting up new virtual resources can cause Service Level Agreement violation. Consequently, predicting cloud resources provisioning gains a lot of attention to scale computational resources in advance. However, most of current approaches do not consider multi-seasonality in cloud workloads. This paper proposes cloud resource provisioning prediction algorithm based on Holt-Winters exponential smoothing method. The proposed algorithm extends Holt-Winters exponential smoothing method to model cloud workload with multi-seasonal cycles. Prediction accuracy of the proposed algorithm has been improved by employing Artificial Bee Colony algorithm to optimize its parameters. Performance of the proposed algorithm has been evaluated and compared with double and triple exponential smoothing methods. Our results have shown that the proposed algorithm outperforms other methods.**

Keywords*—auto-scaling; cloud computing; cloud resource scaling; holt-winters exponential smoothing; resource provisioning; virtualized resources

I. INTRODUCTION

Elasticity feature plays an important role in cloud computing by allowing SaaS providers to allocate and deallocate resources to their running services according to the demand. Elasticity allows SaaS providers to pay only for resources that are used by their cloud services [1]. However, the delay between requesting new resources and it being ready for use violates Service Level Agreement [2]. Therefore, forecasting future resource provisioning is needed to request resources in advance.

Exponential Smoothing is a very popular smoothing method and has been used through years in many forecasting situations [3]. Many researchers have exploited Exponential smoothing methods to predict future resource provisioning for cloud computing applications [4][5]. However, most of them have used double exponential smoothing, which cannot model workloads if there are seasonalities.

Most of cloud-computing applications' workloads are influenced by seasonal factors (e.g., day, week, month, year) and have more than one seasonal pattern [6][7][8]. Workload has intraday seasonal pattern if there is a similarity of request when comparing requests of the corresponding hour from one day to the next day. Intraweek seasonal pattern exists if there is a similarity between requests in two corresponding days from two adjacent weeks [3]. Therefore, there is a strong demand to use predictive approach that is able to capture all seasonality patterns.

This paper proposes resource usage prediction algorithm, which extends Holt-Winters exponential smoothing (HW) method to model multiple seasonal cycles. However, modeling multiple seasonal cycles requires large number of observation values. For example, predicting resource usage with intraday, intra-month, and intra-year seasonality patterns requires at least two years observation values. Moreover, finding optimal parameter values (smoothing constant, trend-smoothing constant and seasonal-smoothing constants) for multiple seasonality model is not an easy task.

Therefore, the proposed algorithm detects seasonality patterns from available historical data by applying seasonality test, and extends HW accordingly to model detected seasonality patterns. While historical data size grows up and more seasonality patterns are detected, HW is gradually extended to be able to model detected seasonality patterns. Furthermore, prediction accuracy of the proposed algorithm has been enhanced by using artificial bee colony algorithm to find near optimal values for its parameters. Thus, unlike most of current resource prediction approaches, the proposed algorithm does not require any minimum number of observations values before applying it. However, good prediction accuracy will not be achieved until several steps have been made.

The proposed algorithm has been evaluated using CloudSim simulator with real Web server log called Saskatchewan Log [6]. Performance of the proposed algorithm has been compared with double and triple exponential smoothing methods. Experimental results have shown that the proposed algorithm outperforms algorithms that use double or triple exponential smoothing methods.

This paper is organized as follows. In Section II related works are overviewed. The proposed algorithm is presented in

Section III. Performance of the proposed algorithm is evaluated in Section IV. Finally, Section V concludes.

II. Related Work

The problem of predicting resource provisioning in cloud computing has been studied extensively over the last few years. Several prediction techniques have been used to predict cloud resource provisioning. However, most of current approaches do not consider multi-seasonality in cloud workload, and most of them use prediction techniques that do not have ability to model more than one seasonal cycle [4][9][10][11].

Islam et al. [12] have proposed framework to predict future resource usage in the cloud. The proposed framework uses two machine-learning algorithms (Neural Network and Linear Regression) with sliding window and cross validation techniques to predict cloud resource usage. The proposed framework is evaluated by using dataset that is collected by using TPC-W benchmark. Statistical metrics is proposed to assess prediction accuracy. However, the proposed framework uses three layers feed-forward Neural Network, which does not able to predict resource utilization when there are long time lags between events. Moreover, the proposed framework is tested with data that are collected from 135 minutes, which does not contain any seasonality. Therefore, prediction with seasonality is not examined.

Kanagala and Sekaran [4] have proposed dynamic threshold-based auto-scaling approach that considers virtual resource start-up and stabilization delays. Virtual resource utilization is predicted by using double exponential smoothing method, thresholds are adapted based on the predicted resource utilization to minimize violation of Service Level Agreement. However, double exponential smoothing method cannot be used to model seasonality.

In [5], Huang et al. have proposed resource utilization prediction model based on double exponential smoothing method. Prediction accuracy of the proposed model has been evaluated using CloudSim simulator, which shows that double exponential smoothing has better prediction accuracy than simple mean based method and weighted moving average method. However, smoothing constant and trend-smoothing constant are determined using trial method, which does not grant quality of the final solution.

Although, seasonal linear regression can be used to predict workload with seasonality, most of current approaches do not consider cloud workload seasonalities and use conventional linear regression to predict cloud resource utilization [1][13][14][15]. In [16], Yang et al. have proposed cost-aware auto-scaling approach, which predicts workload using linear regression model. The problem has been formulated as integer programming problem and solved using greedy heuristic to reduce costs. The proposed approach uses vertical and horizontal scaling methods. Allocated resources are scaled vertically by creating virtual machines on the same cluster node or using unallocated resources available at a particular cluster node to scale up a VM executing on it. Horizontal scaling is used to create virtual machines on other cluster nodes.

To gain benefits from several time series prediction models, Messias et al. [2] have proposed cloud workload prediction methodology that combines several time series forecasting models using genetic algorithm. Each time series prediction model has been assigned a weight, and genetic algorithm adapts the assigned weights to find the best weight combination that maximizes prediction accuracy.

Wei and Blake [17] have proposed an algorithm to predict future resource requirement in the cloud. The proposed algorithm uses five prediction models and differentiates between these models using root-mean-square-error (RMSE). Prediction model with the lowest RMSE is used to predict future resource requirement. Although, the proposed algorithm uses prediction techniques that do not have ability to model seasonality, it can be extended to include more prediction techniques with the ability to model seasonality.

Salah et al. [18] have proposed analytical model based on Markov chains to predict minimal number of VMs and load balancers required to satisfy Service Level Agreement such as throughput and response time. The proposed model has been validated using experimental testsbed deployed on the Amazon Web Services. Discrete-event simulation has been used to verify correctness of the proposed model.

III. Proposed Algorithm

Although many researchers have employed double exponential smoothing for forecasting cloud applications' workload [5][4], double exponential smoothing does not able to model seasonality [3]. HW can be used for forecasting seasonal workloads [3]. However, HW is only able to model workloads with one seasonal pattern and cloud applications' workloads may have more than one seasonal pattern (e.g., intraday, intraweek, intra-month, intra-quarter, intra-year).

Therefore, in this paper, HW has been extended to be able to accommodate multi-seasonal patterns. As shown in equations 1-5, HW has been extended by adding seasonal indices and smoothing equation for each seasonal pattern.

$$S_t = \alpha \frac{X_t}{M(0)} + (1-\alpha)(S_{t-1} + B_{t-1}), \quad 0 < \alpha \leq 1 \qquad (1)$$

$$B_t = \beta(S_t - S_{t-1}) + (1-\beta)B_{t-1}, \quad 0 < \beta \leq 1 \qquad (2)$$

$$I_{i,t} = \gamma_i \frac{X_t}{S_t} \frac{I_{i,t-L_i}}{M(0)} + (1-\gamma_i) I_{i,t-L_i}, \quad 0 < \gamma_i \leq 1 \qquad (3)$$

$$M(k) = \begin{cases} \prod_{i=1}^{n} I_{i,t-L_i+k}, & if\ n \geq 1 \\ 1, & if\ n = 0 \end{cases} \qquad (4)$$

$$\hat{X}_t(k) = (S_t + k\, B_t)\, M(k) \qquad (5)$$

where t is an index denoting a time period, S_t is smoothed value at time t, X_t is observed value at time t, B_t is trend factor at time t, $I_{i,t}$ is seasonal indices for seasonality pattern i, L_i is number of periods in a completed seasonal cycle for seasonality pattern i, α is the smoothing constant, β *is* the trend-smoothing constant, γ_i is seasonal-smoothing constant for seasonality pattern i, n is number of seasonality patterns, and $\hat{X}_t(k)$ is the k-step-ahead forecast at time t.

Initial smoothed value for the level, S_1 , is calculated as average of the first $2L_1$ periods, which are periods in the first two cycles from the first seasonal pattern. If there is no seasonality patterns, S_1 is initialized by the first observation value X_1. Initial value for the trend factor, T_1, is calculated as $1/L_1$ of average of the difference between first L_1 observations and second L_1 observations. If there is no seasonality patterns, T_1 is initialized as a difference between second observation value and first observation value ($X_2 - X_1$).

For each seasonal cycle, at least three completed seasonal data are required to initialize its seasonal indices. Seasonal indices are initialized as average of ratios of observed value to its centered moving average (calculated from L_i periods around observed value), taken from the corresponding period in each of the first two completed seasonal data, which starts from $t = L_i/2$. For example, seasonal indices for seasonality pattern i are calculated as following:

$$I_{i,t} = \left(\frac{X_t}{A_t} + \frac{X_{t+s_1}}{A_{t+s_1}}\right)/2 \ , t = \frac{L_i}{2}, \frac{L_i}{2} + 1 \ , \frac{L_i}{2} + 2 \ , \dots , \frac{L_i}{2} + L_i$$

where A_j is centered moving average around X_j for L_i periods

$$A_j = \sum_{i=j-L_i/2}^{j-1+L_i/2} X_i \Big/ L_i$$

Finally, artificial Bee colony algorithm is applied to determine near optimal values for smoothing constant, trend-smoothing constant and seasonal-smoothing constants that minimize Mean Squared Error (MSE).

$$\mathrm{MSE_t} = \frac{1}{\mathrm{t}} \sum_{\mathrm{i=1}}^{\mathrm{t}} (\hat{X}_i(k) - X_{i+k})^2$$

Algorithm 1 shows steps of the proposed algorithm. The first input is initial list of observation values that contains 60 observation values (from 60 minutes). This number of observation values is specified to start with enhanced accuracy. The second input is the list of completed seasonal cycles' length of expected seasonal patterns. Instead of applying seasonality test periodically, the second input specifies time points to test existence of seasonality patterns. The outputs are list of predicted values and list of seasonal cycles' length of detected seasonal patterns.

In the first line, initial smoothed value s_1 is set to the observed value x_1, and initial trend factor b_1 is set to $(x_2 - x_1)$. Best values for smoothing constant α and trend-smoothing constant β are obtained by using Bee Colony Algorithm (Algorithm 2). At this point, number of seasonal cycles $n = 0$. Therefore, equations 1-5 are minimized to the following equations, which represent equations associated with Double Exponential Smoothing.

$$S_t = \alpha \ X_t + \ (1 - \alpha \ )(S_{t-1} + B_{t-1}), \ \ 0 < \alpha \leq 1 \qquad (6)$$

$$B_t = \ \beta (S_t - S_{t-1}) + (1 - \beta) B_{t-1} \ , \ \ 0 < \beta \leq 1 \qquad (7)$$

$$\hat{X}_t(k) = S_t \ + \ k \ B_t \qquad (8)$$

Therefore, prediction accuracy of the proposed algorithm during the interval from $t = 61 \ to \ t = 3 * L_1 - 1$ (where L_1 is the number of periods in completed seasonal cycle for the first seasonal pattern) is very similar to prediction accuracy of double exponential smoothing.

If t equals to $3 * l_i^{'}$, where $l_i^{'} \in L^{'}$, seasonality test is applied to check if the list of observed values X has seasonal pattern with length $l_i^{'}$ or not. New seasonal pattern is detected if autocorrelation coefficient is greater than or equal 0.3. Length of the detected seasonal pattern $l_i^{'}$ is added to the list of seasonal cycles' length L, and number of detected seasonal patterns n is increased. List of seasonal indices for the new seasonal pattern is calculated and added to I. Smoothing constant, trend-smoothing constant, and seasonal-smoothing constants are updated to the near optimal values using artificial Bee colony algorithm. Finally, extended formula is employed to predict future required resources.

ALGORITHM 1: The proposed algorithm

INPUTS:
X: initial list of observation values
L': list of expected seasonal cycles' length

OUTPUTS:
$\hat{X}(k)$: list of k-step-ahead predicted values
L: list of seasonal cycles' length

Begin
$s_1 = x_1$
Initialize S and add s_1 to S
$b_1 = x_2 - x_1$
Initialize trend factor list B and add b_1 to B
Get Best Constants Using Bee Colony Algorithm
$n = 0$, where n is the number of seasonal cycles in X
$t = 1$, where t is an index denoting a time period
while $t \leq 60$
  Calculate s_t using equation 1 and add it to S
  Calculate b_t using equation 2 and add it to B
  Calculate $\hat{X}_t(k)$ using equation 5 and add it to $\hat{X}(k)$
  $t++$
end while
for each new observation value x_t at time t
  Add $x_t \ to \ X$
  if $t/3 \in L'$
    Apply seasonality test
    if autocorrelation coefficient≥ 0.3
      $n++$
      Add $t/3$ to L
      Initialize seasonal indices list I_n for seasonal cycle with length t/3
      Get Best Constants Using Bee Colony Algorithm
    end if
  end if
  Calculate s_t using equation 1 and add it to S
  Calculate b_t using equation 2 and add it to B
  Calculate $I_{i,t}$ using equation 3 for all $i = 1,2,..,n$ and add it to I
  Calculate $\hat{X}_t(k)$ using equation 5 and add it to $\hat{X}(k)$
end for
return
End

Algorithm 2 shows steps of determining best values for smoothing constant α, trend-smoothing constant β , and

seasonal-smoothing constants γ by using artificial Bee colony optimization algorithm.

At the beginning, initial population P is initialized with ns scout bees, which are randomly scattered across solution space. Here, all constants (smoothing constant α, trend-smoothing constant β, and seasonal-smoothing constants γ) are greater than zero and less than or equal one. For each scout bee in ns, flower patch is delimited that contains its neighborhood.

MSE is calculated for each scout bee by applying equations 1-5 from $t = l_n$ to $t = \|X\|$, where l_n is number of periods in completed seasonal cycle for the largest seasonal pattern. if $n = 0$, MSE is calculated by applying equations 1-5 from $t = 2$ to $t = \|X\|$. Scouts are sorted in ascending order according to their MSE.

Best sites nb with lowest MSE are selected from ns, and elite sites ne with most lowest MSE are selected from nb.

Each scout in nb performs waggle dance to recruit forager bees to search further in its flower patch. Such that, number of recruited forager bees to ne (nre) is greater than number of recruited forager bees to the remaining best sites $nb - ne$ (nrb).

ALGORITHM 2: Determine Best Constants Using Bee Colony Algorithm

INPUTS:
S: list of smoothed values
X: list of observed values
B: trend factor list
I: seasonal indices list
n: number of detected seasonality patterns
L: list of seasonal cycles' length
$MaxIter$: maximum iteration number
$MaxError$: maximum allowed error

OUTPUTS:
Near optimal values for smoothing constant α, trend-smoothing constant β, and seasonal-smoothing constants γ

Begin
1: Generate initial population P with ns scout bees
2: Specify flower patch for each scout in nb
3: Calculate MSE for each scout in P
4: Sort scouts in P in ascending order based on their MSE values
5: $i = 0$
6: **while** $i \leq MaxIter$ or $FitnessValue_i - FitnessValue_{i-1} \leq MaxError$
7: $i++$
8: Select best sites nb from ns
9: Select elite sites ne from nb
10: Recruit forager bees to ne and $nb - ne$
11: Apply local search to find fittest bee of each flower patch
12: Generate random solutions for non-best sites $ns - nb$
13: Calculate MSE for non-best sites ns − nb
14: Sort all scouts in ns in ascending order based on their MSE values
15: $FitnessValue_i = MSE$ of the first scout in the sorted ns
16: **end while**
17: Determine constants' value according to fittest scout in population P
18: **return**
End

To find fittest bee of each flower patch, recruited forager bees are randomly distributed in flower patch. MSE is calculated for each bee. If there is recruited forager bee with MSE lower than MSE of its scout bee, fittest bee will be selected as a new scout. Otherwise, flower patch will be shrunken around its scout. After pre-specified number of search cycles, the fittest bee of each flower patch is returned as a local optimal solution.

New solutions are generated randomly for non-best sites $ns - nb$, and all scout in ns are sorted in ascending order according to their MSE. This search cycle will be repeated until reaching termination condition. Finally, values of smoothing constant α, trend-smoothing constant β, and seasonal-smoothing constants γ are obtained from fittest scout bee in current population.

IV. Performance Evaluation

To evaluate performance of the proposed algorithm, its performance have been compared with double and triple exponential smoothing methods. The following subsections, describe evaluation environment settings and discuss simulations' results.

A. Evaluation environment settings

The proposed algorithm has been evaluated using real Web server log called Saskatchewan Log [6]. Saskatchewan log contains HTTP requests to the University of Saskatchewan's WWW server, which is located in Saskatoon, Saskatchewan, Canada. This log was collected from 00:00:00 June 1, 1995 to 23:59:59 December 31, 1995, a total of 214 days [6].

Cloudlets have been generated according to Saskatchewan log and sent to CloudSim simulator. For each minute, CloudSim simulator calculates total required CPU to process incoming requests without violating Service Level Agreement. The proposed algorithm receives required CPU as observed value and predicts required CPU after *k*-minutes. *K* has been set to 15, where *k* is a virtual machine startup delay.

To evaluate accuracy of the proposed algorithm, three evaluation metrics have been used:

- *Mean absolute percentage error (MAPE)*, which is defined as following:

$$MAPE_t = \frac{1}{t}\sum_{i=1}^{t}\frac{\left|\hat{X}_t(k) - X_{t+k}\right|}{X_{t+k}}$$

where $MAPE_t$ is mean absolute percentage error at time t, $\hat{X}_t(k)$ is the *k*-step-ahead forecast at time t, and X_{t+k} is observed value at time $t + k$. A smaller value of $MAPE_t$ implies a better prediction accuracy.

-*Percentage of predictions within 25%* (*PRED(25)*), percentage of prediction within 25% at time t is defined as following:

$$PRED(25)_t = \frac{1}{t}\left\|\left\{\hat{X}_i(k): \frac{\left|\hat{X}_i(k) - X_{i+k}\right|}{X_{i+k}} < 25\%,\quad 0 \leq i \leq t\right\}\right\|$$

$PRED(25)_t$ values are between 0 and 1. Prediction will be more effective if $PRED(25)_t$ value is closer to 1.

-*Root Mean Squared Error* (RMSE), RMSE at time t is defined as following:

$$\mathrm{RMSE_t} = \sqrt{\frac{1}{t}\sum_{i=1}^{t}(\hat{X}_i(k) - X_{i+k})^2}$$

A smaller value of $\mathrm{RMSE_t}$ implies better prediction accuracy.

B. Evaluation results

Although, the proposed algorithm has been evaluated with many real workload traces such as [6][7][8] in this evaluation only one of them has been shown, which is Saskatchewan-http.

Fig. 1 compares the proposed multi-seasonal algorithm with double and triple exponential smoothing methods using Mean Absolute Percentage Error (MAPE), which has been defined in the previous section. As shown in Fig. 1, MAPE of the proposed multi-seasonal algorithm stays below 29% while triple and double are above 44% and 135% respectively. Fig. 2 shows that more than 57% of predicted values by using the proposed multi-seasonal algorithm are with prediction error less than 25%. In another side, 38% of triple exponential smoothing predictions are within 25%, and 8-18% of double exponential smoothing predictions are within 25%. Finally, Root Main Square Error of the proposed multi-seasonal algorithm has been compared with double and triple exponential smoothing methods in Fig. 3, which shows that RMSE of the proposed multi-seasonal algorithm is better than other methods.

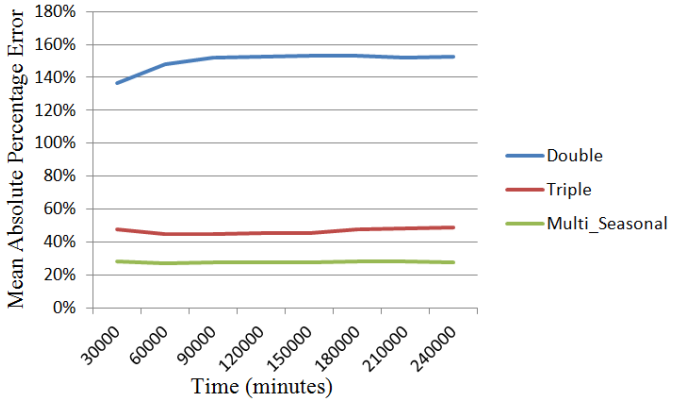


Fig. 1. Mean Absolute Percentage Error comparison

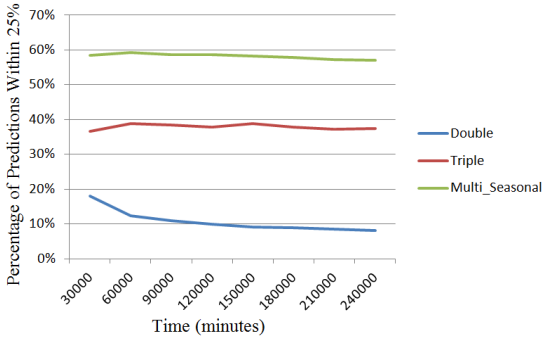


Fig. 2. Percentage of Predictions Within 25% comparison

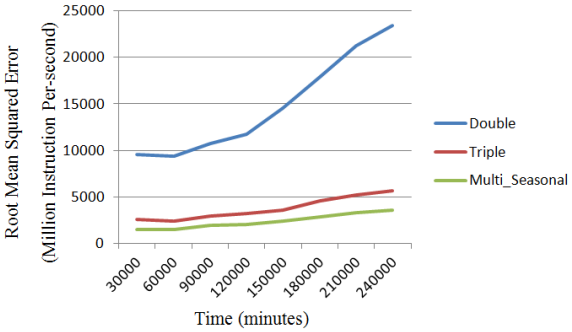


Fig. 3. Root Main Square Error comparison

V. Conclusion

This paper has proposed predictive algorithm to predict cloud resource provisioning. According to available historical data and detected seasonal cycles, Holt-Winters exponential smoothing method has been extended to allow modeling multiple seasonal cycles with minimum number of observation values. Artificial Bee Colony algorithm has been exploited to find near optimal parameters value for the proposed algorithm. Prediction accuracy of the proposed algorithm has been evaluated by using CloudSim simulator with real workload called Saskatchewan-http. Our results have shown the effectiveness of the proposed algorithm among other methods. Finally, the paper concludes that modeling multiple seasonal cycles during predicting cloud resource provisioning is an essential step toward accurate cloud resource prediction.

As future work, long short-term memory recurrent neural networks will be incorporated with the proposed algorithm to predict cloud resource utilization when there are very long and variant time lags between events. Because, in seasonality patterns, seasonal cycle length is considered constant for each seasonal pattern. However, in some cases, lags between events are variant and have to be considered during prediction.

REFERENCES

[1] M. Ghobaei-Arani, S. Jabbehdari, and M. A. Pourmina, "An autonomic approach for resource provisioning of cloud services," Cluster Computing, vol. 19, no. 3, pp. 1017–1036, 2016. DOI: 10.1007/s10586-016-0574-9

[2] V. R. Messias, J. C. Estrella, R. Ehlers, M. J. Santana, R. C. Santana, and S. Reiff-Marganiec, "Combining time series prediction models using genetic algorithm to autoscaling web applications hosted in the cloud infrastructure," Neural Computing and Applications, pp. 1–24, 2015. DOI: 10.1007/s00521-015-2133-3

[3] J. W. Taylor, "Exponentially weighted methods for forecasting intraday time series with multiple seasonal cycles," International Journal of Forecasting, vol. 26, no. 4, pp. 627 – 646, 2010. DOI: http://dx.doi.org/10.1016/j.ijforecast.2010.02.009

[4] K. Kanagala and K. Sekaran, "An approach for dynamic scaling of resources in enterprise cloud," in 2013 IEEE 5th International Conference on Cloud Computing Technology and Science (CloudCom), vol. 2, Dec 2013, pp. 345–348. DOI: 10.1109/CloudCom.2013.167

[5] J. Huang, C. Li, and J. Yu, "Resource prediction based on double exponential smoothing in cloud computing," in 2nd International Conference on Consumer Electronics, Communications and Networks (CECNet), 2012, April 2012, pp. 2056–2060. DOI: 10.1109/CECNet.2012.6201461

[6] Clarknet-http, Two weeks of http logs from the Clarknet WWW server. Metro Baltimore-Washington DC area, USA. [online] http://ita.ee.lbl.gov/html/contrib/ClarkNet-HTTP.html (Accessed on October 1, 2016)

[7] Nasa-http, Two months of http logs from NASA Kennedy Space Center WWW server in Florida, USA. [online] http://ita.ee.lbl.gov/html/contrib/NASA-HTTP.html (Accessed on October 1, 2016)

[8] Saskatchewan-http, Seven months of http logs from the Saskatchewan's WWW server. Saskatchewan University, Saskatchewan, Canada. [online] http://ita.ee.lbl.gov/html/contrib/Sask-HTTP.html (Accessed on October 1, 2016)

[9] S. Ajila and A. Bankole, "Cloud client prediction models using machine learning techniques," in 2013 IEEE 37th Annual Computer Software and Applications Conference (COMPSAC), July 2013, pp. 134–142. DOI: 10.1109/COMPSAC.2013.21

[10] Z. Shen, S. Subbiah, X. Gu, and J. Wilkes, "Cloudscale: Elastic resource scaling for multi-tenant cloud systems," in Proceedings of the 2Nd ACM Symposium on Cloud Computing, ser. SOCC '11. New York, NY, USA: ACM, 2011, pp. 5:1–5:14. DOI: 10.1145/2038916.2038921

[11] I. K. Kim, J. Steele, Y. Qi, and M. Humphrey, "Comprehensive elastic resource management to ensure predictable performance for scientific applications on public iaas clouds," in Utility and Cloud Computing (UCC), 2014 IEEE/ACM 7th International Conference on, Dec 2014, pp. 355–362. DOI: 10.1109/UCC.2014.45

[12] S. Islam, J. Keung, K. Lee, and A. Liu, "Empirical prediction models for adaptive resource provisioning in the cloud," Future Gener. Comput. Syst., vol. 28, no. 1, pp. 155–162, Jan. 2012. DOI: 10.1016/j.future.2011.05.027

[13] J. Yang, C. Liu, Y. Shang, Z. Mao, and J. Chen, "Workload predicting-based automatic scaling in service clouds," in 2013 IEEE Sixth International Conference on Cloud Computing, June 2013, pp. 810–815. DOI: 10.1109/CLOUD.2013.146

[14] A. Biswas, S. Majumdar, B. Nandy, and A. El-Haraki, "Automatic resource provisioning: A machine learning based proactive approach," in 2014 IEEE 6th International Conference on Cloud Computing Technology and Science (CloudCom), Dec 2014, pp. 168–173. DOI: 10.1109/CloudCom.2014.147

[15] A. Bankole and S. Ajila, "Predicting cloud resource provisioning using machine learning techniques," in 2013 26th Annual IEEE Canadian Conference on Electrical and Computer Engineering (CCECE), May 2013, pp. 1–4. DOI: 10.1109/CCECE.2013.6567848

[16] J. Yang, C. Liu, Y. Shang, B. Cheng, Z. Mao, C. Liu, L. Niu, and J. Chen, "A cost-aware auto-scaling approach using the workload prediction in service clouds," Information Systems Frontiers, vol. 16, no. 1, pp. 7–18, 2014. DOI: 10.1007/s10796-013-9459-0

[17] Y. Wei and M. B. Blake, "Proactive virtualized resource management for service workflows in the cloud," Computing, vol. 98, no. 5, pp. 523–538, 2016. DOI: 10.1007/s00607-014-0419-4

[18] K. Salah, K. Elbadawi, and R. Boutaba, "An analytical model for estimating cloud resources of elastic services," Journal of Network and Systems Management, vol. 24, no. 2, pp. 285–308, 2016. DOI: 10.1007/s10922-015-9352-x

Optimal Path Planning using RRT* based Approaches: A Survey and Future Directions

Iram Noreen
Department of Computer Science
COMSATS Institute of Information Technology
Lahore, Pakistan

Amna Khan
Department of Computer Science
COMSATS Institute of Information Technology
Lahore, Pakistan

Zulfiqar Habib
Department of Computer Science
COMSATS Institute of Information Technology
Lahore, Pakistan

***Abstract*—Optimal path planning refers to find the collision free, shortest, and smooth route between start and goal positions. This task is essential in many robotic applications such as autonomous car, surveillance operations, agricultural robots, planetary and space exploration missions. Rapidly-exploring Random Tree Star (RRT*) is a renowned sampling based planning approach. It has gained immense popularity due to its support for high dimensional complex problems. A significant body of research has addressed the problem of optimal path planning for mobile robots using RRT* based approaches. However, no updated survey on RRT* based approaches is available. Considering the rapid pace of development in this field, this paper presents a comprehensive review of RRT* based path planning approaches. Current issues relevant to noticeable advancements in the field are investigated and whole discussion is concluded with challenges and future research directions.**

Keywords*—optimal path; mobile robots; RRT*; sampling based planning; survey; future directions

I. INTRODUCTION

The term path planning refers to collision free path generation from an initial state to a specified goal state with optimal or near optimal cost. Considering different applications and constraints of robots, optimal criteria could be based on one or more conditions such as shortest physical distance, smoothness, low risk, less fuel requirements, maximum area coverage, and low energy consumption. Hence, in perspective of path planning for mobile robots optimal path refers to find a feasible plan with optimized performance according to application specified criterion [1]. Optimal path planning is also influenced by the holonomic and non-holonomic constraints. According to LaValle, the term non-holonomic refers to the differential constraints (restrictions on permissible velocities) that are not completely integrable, such as car-like robots and the others are holonomic constraints such as robotic arm [1].

Path planning algorithms are of vital importance for motion planning of mobile robots due to their numerous applications in autonomous cars [2], Unmanned Aerial Vehicles (UAVs) [3], forklifts [4], surveillance operations [5], medical [6], planetary and space missions [1, 7]. Initial complete practical planners such as Road Map (RM), Potential Fields, and Cell Decomposition (CD) techniques are unable to deal with dynamic and complex high dimension problems [1, 7-10]. Computational complexity of complete planners limits their applications to low dimensional problems [11].

Grid based algorithms such as Dijkstra [12], wavefront [13] , A* [14], D* [15], and Phi* [16] are resolution-complete and are computationally expensive for high dimensional complex problems. Evolutionary algorithms such as Particle Swarm Optimization (PSO) [17-19], Ant Colony Optimization (ACO) [20] and Genetic Algorithm (GA) [21] are suitable for multi-objective problems. Many other evolutionary algorithms such as Artificial Bee Colony (ABC) [22], Bacterial Foraging Optimization (BFO) [23], Bio Inspired Neural Networks [24, 25], and Fire Fly algorithm [26] are often trapped in local optimum, and bear high computational cost. Moreover, they are highly sensitive to search space size and data representation scheme of problem [27, 28].

Sampling Based Planning (SBP) approaches are the most influential advancement in path planning [7, 8]. Major advantages of Sampling Based Planning (SBP) are low computational cost, applicability to high dimensional problems and better success rate for complex problems [8, 29]. SBPs are probabilistic complete, i.e., it finds a solution, if one exists, provided with infinite run time [4, 8]. Most popular SBP algorithms are Probabilistic Roadmap (PRM) [7, 8, 30], Rapidly-exploring Random Tree (RRT) [11, 31] and Rapidly-exploring Random Tree Star (RRT*) [7]. PRM based methods [7, 32] are mostly used in highly structured static environment such as factory floors [11, 29, 33]. They are well suited for holonomic robots but could be extended for non-holonomic as well [31]. On the other hand RRT and RRT* based approaches [7] naturally extend non-holonomic constraints [11] and support dynamic environment as well.

Introduced by Karaman and Frazzoli [7], RRT* was a major breakthrough in optimal path planning for high dimensional problems. RRT* has proven asymptotically optimal property, i.e., RRT* always converges to an optimal solution, if adequate run time is provided. RRT* has gained tremendous success in solving high dimensional complex problems with numerous successful applications. A survey on Sampling Based Planning (SBP) approaches for mobile robots was presented in [8]. However, considerable body of research has specifically addressed the problem of optimal path planning focusing RRT* in recent years as compared to other SBP approaches. Rapid pace of development in optimal path planning using RRT* based approaches has grown it into a

family of algorithms [2, 3, 28, 33-50]. To the best of our knowledge, no updated survey exists on RRT* based approaches. This paper is an effort to review the major breakthroughs in RRT* based approaches providing link to the most successful works in the field. Moreover, current state of the art is surveyed to explore recent contributions and future directions in optimal path planning.

This paper has been organized in such a way that a discussion on RRT* methodology is presented in next section. Section III categorizes RRT* based path planning approaches in recent years. State of the art techniques are summarized in Section IV. Section V presents challenges followed by conclusion along with future recommendations in Section VI.

II. RRT* Methodology

This section introduces important path planning concepts related to RRT* in order to provide a better understanding of this study. It is essential to introduce the basic operations of RRT* prior to describe its variant approaches. These procedures are found in all RRT* variants, but their implementation may differ in different planners and applications.

A. Problem Formulation

RRT* based approaches operate in the configuration space. This configuration space is a set of all possible transformations which are applicable to the robot. [1, 51]. Let the given configuration space be denoted by a set $Z \subset \mathbf{R}^n, n \in \mathbf{N}$ where n represents the dimension of the given space and $\mathbf{N}$ is a set of positive integers. Configuration space occupied by obstacles is denoted by $Z_{obs} \subset Z$ and obstacle-free region is denoted by $Z_{free} = Z / Z_{obs}$. $z_{goal} \subset Z_{free}$ is the goal and $z_{init} \subset Z_{free}$ is the starting point. z_{init} and z_{goal} are provided to planner as input. The problem is to find an optimal collision free path between initial z_{init} and goal z_{goal} states in Z_{free}, with minimum path cost in the least possible time $t \in \mathbf{R}$, where $\mathbf{R}$ is the set of real numbers.

B. Tree Expansion in RRT*

RRT* constructs multiple short paths randomly organized as tree instead of one long path. It originates tree from initial state z_{init} to find a path towards goal state z_{goal}. The tree gradually improves with iterations. In each iteration, a sampling process selects a random state say z_{rand} from configuration space Z . The random sample z_{rand} is rejected if it lies in Z_{obs}. However, if it lies in Z_{free} then a nearest node say $z_{nearest}$ is searched in tree T according to a defined metric ρ . If z_{rand} lies in Z_{free} and is also accessible to $z_{nearest}$ according to predefined step size, then a local planner inserts it in tree by connecting z_{rand} and $z_{nearest}$. Otherwise, planner returns a new node z_{new} by using a steering function and adds it in tree by connecting it with $z_{nearest}$. This property of RRT*, to explore region in Z_{free} is called Voronoi bias. A collision checking process is performed to ensure collision free connection between z_{new} and $z_{nearest}$ .The Node expansion process is illustrated in Fig. 1.

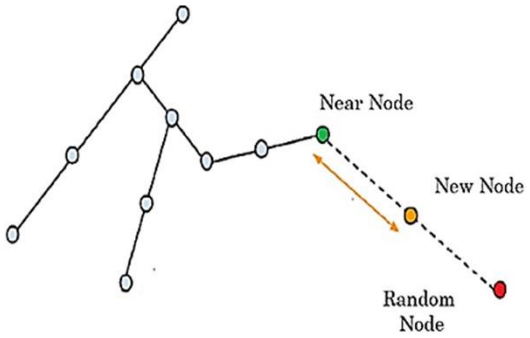


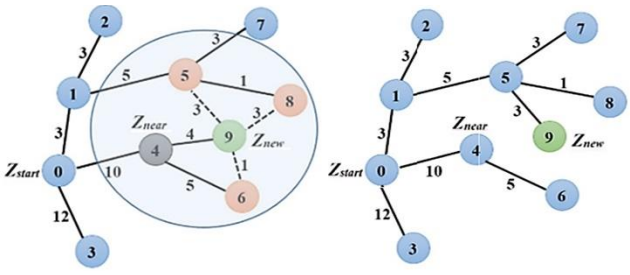
Fig. 1. RRT* Tree expansion process [52]

If z_{new} is found collision free then near neighbors of z_{new} are searched within the area of a ball of radius defined by

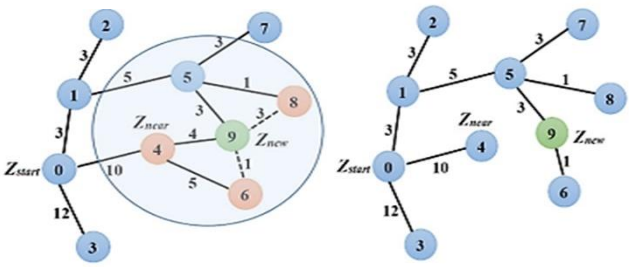
$$k = \gamma(\log(n)/n)^{1/d} \text{ [7]}, \quad (1)$$

where *d* is the configuration space dimension and γ is the planning constant based on environment. Within the area defined by (1), neighbor z_{min} with least cost is selected to be parent of z_{new}. Procedure of near neighbor search is similar to *k*-near neighbor problem to find out the best parent node z_{min} of new node z_{new} before its insertion in tree. New node z_{new} is inserted as child of z_{min} in tree. Further, the cost of near neighbor's parent node is also compared with the cost of z_{new} . If z_{new} gives less cost as parent, then rewiring process rebuilds the tree for minimum parent cost within the area identified by (1) [53]. This process is shown in Fig. 2.

The process of selecting least cost parent and rewiring tree are two most promising features of RRT* and contribute to asymptotic optimal property of RRT* [7]. Though best parent selection and rewiring of tree improve the path quality. However, these features have an efficiency trade-off with path quality and make convergence slow as number of nodes in the tree increase. When z_{goal} is found, a path connecting z_{init} and z_{goal} is established. This path is improved as planner continues until a predefined number of iterations are executed or given time expires. The RRT* Algorithm is described as Algorithm 1.



(a) Find near neighbors (b) Select best parent



(c) Check cost for rewiring (d) Rewired tree

Fig. 2. Near neighbor search and rewiring operations in RRT* [52]

Algorithm 1. T = (V, E) ← RRT*(z_{init})

```
T ← InitializeTree();
T ← InsertNode(Ø, z_init, T);
for i=0 to i=N do
    z_rand ← Sample(i);
    z_nearest ← Nearest(T, z_rand);
    (z_new, U_new) ← Steer (z_nearest, z_rand);
    if Obstaclefree(z_new) then
        z_near ← Near(T, z_news, |V|);
        z_min ← Chooseparent (z_near, z_nearest, z_new);
        T ← InsertNode(z_min, z_new, T);
        T ← Rewire (T, z_near, z_min, z_new);
return T
```

III. Methodologies based on RRT* Algorithm

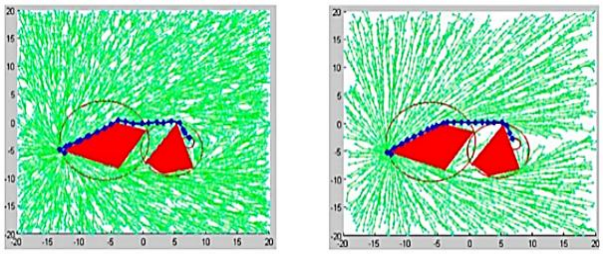
This section provides review of optimal path planning using RRT* in recent six years with major breakthroughs in the field. To provide a better understanding of research body, we have classified these approaches based on the similar concepts such as type of environment information available, the structure of the tree and the constraints managed by approach. RRT* have been used in online mode or offline mode depending upon availability of environment information. If environment parameters are unknown or highly uncertain then local planning is performed, also called *online* (sensor based, or reactive). Whereas, a known environment requires global planning, also called *offline* (map based) [23, 54]. Further, RRT* variants based on bidirectional trees also exist in literature, which generate two trees simultaneously from start and goal states. Further, RRT* based approaches considering *non-holonomic* constraints also exist.

A. Single Directional Holonomic RRT* Approaches

This section presents RRT* based approaches which generate path for holonomic robots and construct a single tree originating from initial state z_{init} towards goal state z_{goal} to find path in search space. Both online and offline approaches are discussed in this subsection. S. Karaman et al. [4] presented an online Anytime variant of RRT*. Basic idea of Anytime RRT* is to deal with the issue of large computational time by executing the planner for a predefined planning time. Once an initial path is obtained and stored, then rest of the time is used to improve initial solution [55, 56]. Anytime RRT* introduced two key features called committed trajectories and branch-and-bound adaptation. Strategy of committed trajectory originates robot's movement to follow first segment of initially planned path while improving remaining segments of the path using iterative strategy. Whereas Branch-and-bound optimizes the tree for optimal cost. Anytime RRT* improved trajectory and computational efficiency in simulation and in real-time implementation using forklift robot.

Despite success stories, RRT* was suffering from high memory consumption due to large expansion of its search space. Adiyatov and Varol [33] introduced memory efficient version of RRT*, called RRT* Fixed Nodes (RRT*FN). RRT*FN allows limited number of nodes in tree. When tree is expanded to a preset fixed number of nodes then new node can only be inserted by deleting the old node.

The old nodes are deleted according to a defined node removal policy. RRT*FN implies a global node removal procedure and a local node removal procedure for this purpose. Local node removal procedure deletes nodes with single child from near neighbors during rewiring operation, if new node has better cumulative path cost for their child node as parent. In case when no such nodes are identified during rewiring operation then a global node removal procedure is used, which searches entire tree to find nodes without children and deletes them. When both local and global schemes could not found such nodes then new node is not inserted [33]. Hence, search space of RRT*FN consumes less memory by forcing a fixed number of nodes in the tree. Difference of tree density between RRT* and RRT*FN is evident from Fig. 3 (a) and Fig. 3 (b). Such memory efficient versions of RRT* are useful in robots and embedded systems with limited memory [33].



(a) RRT* (b) RRT*FN

Fig. 3. Effect of fixed nodes after 5000 iterations in tree

Nasir et al. [35] presented an offline variant of RRT* called RRT*-Smart to address the issue of slow convergence. RRT*-Smart introduced two major features called intelligent sampling and path optimization. Initial path finding procedure in RRT*-Smart is similar to RRT*. However, once a path is found, it is optimized based on triangular inequality principle to remove redundant nodes [35]. Optimization task generates beacon nodes to further improve path cost. After optimization, it uses both intelligent and uniform sampling strategies alternatively according to defined *Biasing Ratio* for the rest of the iterations using

$$Biasing\ Ratio = (n / Z_{free}) * B \quad [35], \qquad (2)$$

where B is a programming constant and n is total number of nodes in tree. Intelligent sampling is biased towards beacon nodes. Each time it gets a new path with shorter cost, it optimizes the path again and identifies new beacon nodes. This process is based on a Biasing Radius to set radius for

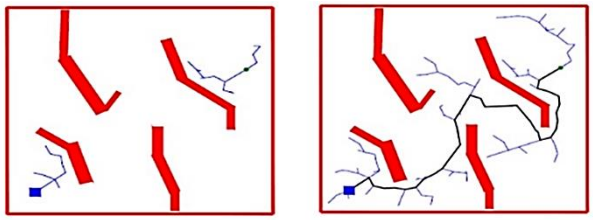
intelligent exploration around selected beacons. The proposed approach accelerated the convergence rate with improved path cost and time efficiency. However, intelligent sampling has a trade-off between rate of convergence and rate of exploration. Therefore, frequency of intelligent sampling needs careful adjustment according to different environment types. An experimental comparison for performance evaluation is also provided in [52] for RRT, RRT* and RRT*-Smart.

Another offline approach called Informed RRT* was presented by Gammell et al. [34] for optimal path planning in narrow passages. They proposed a direct subset sampling technique for configuration space exploration. Once an initial path is found, it further explores configuration space within a limited elliptical area defined by an ellipsoidal informed subset. As area of the ellipse decreases, it also improves the rate of convergence and path quality.

Arslan and Tsiotras [50] proposed RRT* variant called RRT$^{\#}$ (RRT "sharp") to address the issue of slow convergence. RRT$^{\#}$ used two processes during each iteration namely exploration and exploitation. Exploration performs the extension process whereas exploitation uses a global re-planning procedure to keep track of promising nodes of tree. Promising nodes are the ones which are good candidate to contribute in the final path with lowest cost. During each iteration, RRT$^{\#}$ updates information about promising nodes and prioritize them for re-planning in next iteration. Hence, it makes fast convergence by expanding promising nodes towards goal region and exploiting available node information to the highest degree at each iteration.

B. Bidirectional Holonomic RRT Approaches*

All approaches discussed above build single tree in configuration space. This section gives insight on bidirectional approaches. Bidirectional approaches generate two trees simultaneously from start and goal states directing towards each other, as shown in Fig. 4 (a) and 4 (b). Use of bidirectional tree was initially proposed by Kuffner and LaValle in RRT-Connect [11]. They used it initially for motion planning of 7-DOF arm of animated characters used in 3D virtual world. Hence, it was specifically designed for path planning problems with no differential constraints [11].



(a) Two trees growing from start and goal (b) Joined trees

Fig. 4. RRT-Connect, Growing two trees towards each other [11]

Moreover, RRT-connect is not asymptotically optimal like RRT*. Applying bidirectional trees to asymptotically optimal RRT* requires neighborhood rewiring in two trees resulting in high computational cost. Though, bidirectional approaches execute faster for holonomic robots. However, when used for non-holonomic problems, they made very slow convergence. This is due to the fact that managing non-holonomic constraints using connect heuristic of bidirectional tree does not guarantee the connection of both trees [57]. Therefore bidirectional variants of RRT* are considered suitable only for holonomic robots [57, 58].

An asymptotic optimal variant of RRT* [4] and RRT-connect [11] called Bidirectional RRT* (B-RRT*) was proposed by Akgun and Stilman [58]. It showed empirical results indicating fast convergence and path refinement using sample rejection with an admissible heuristic. Though, this procedure selects only promising nodes but also affects space exploration. Moreover, attempt to connect both trees in each iteration incurred computational overhead.

Another bidirectional RRT* was presented by Jordan and Perez [36] for optimal path planning also called Optimal B-RRT*. It was provably asymptotically optimal bidirectional approach with improved convergence rate using a number of heuristic techniques [36]. However, use of multiple heuristics also increased computational overload. Moreover, these biased heuristics interfered with the algorithm characteristics (such as exploration, node rejection, cost function) and limited its application.

Another approach called Intelligent Bidirectional RRT* (IB-RRT*) [37] was proposed by Qureshi and Ayaz for complex cluttered environment. It used an intelligent sample insertion heuristic technique. Simulation results of IB-RRT* showed fast convergence towards optimal path using less memory resources in comparison with RRT* and Bi-RRT.

Recently, Yi et al. [45] presented Homotopy-Aware RRT* (HARRT*) based on bidirectional RRT* [58]. HARRT* is inspired with the idea of homotopy, i.e., to plan path from one topological space to another by human intervention. This approach addresses the planning problems of human-robot team interactions in search and rescue, police, and military operations. Effectiveness of proposed approach is theoretically proved using case studies. However, further investigation of the proposed approach remains to explore using simulations and real world experiments.

C. Non-holonomic and Kinodynamic RRT Approaches*

Non-holonomic robots are car-like robots which have to perform complex motions to achieve a particular direction. This phenomenon also restricts geometry of the path [1]. A car-like robot needs to change its position coordinates in order to rotate around its axis (see Fig. 5). It is under-actuated due to the non-holonomic constraints imposed by the wheels. Therefore, non-holonomic path planning requires satisfying both internal constraints (physical limitations of robot) and external constraints (obstacles in environment) [1, 38, 59]. Further, if kinodynamic constraint arises, then it also affects path planning mechanism.

Kinodynamic planning [60] refers to motion planning problems for which velocity and acceleration bounds must be satisfied. Precisely, kinodynamic is an umbrella term used to deal with kinematics (position, bounds on velocity and acceleration) and dynamic constraints (force) simultaneously [1, 60]. Non-holonomic planning deals with either both (i.e., kinodynamic) constraints or kinematic constraints only [1, 60].

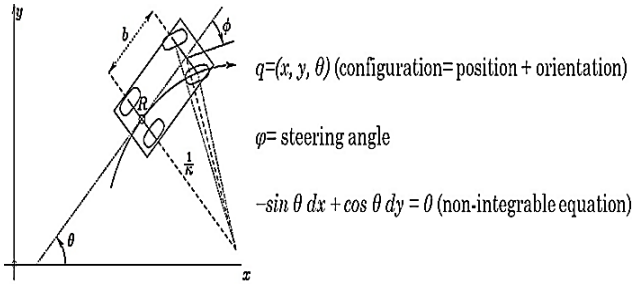


Fig. 5. Non-holonomic constraints of car-like robot [61]

As RRT* connects pair of states using straight lines, which is not feasible for kinodynamic systems due to the differential constraints. Prior kinodynamic extensions of RRT* such as Kinodynamic-RRT* [62] and LQR-RRT* [63] only satisfy bounded sub optimality and require RRT* to re-propagate the tree partially during each iteration. Thus, making these approaches computationally expensive.

In recent years, many RRT* based planners were focused to solve the optimal path planning problem for car-like robots dealing with non-holonomic or kinodynamic constraints. Webb and Berg [49, 64] presented a kinodynamic extension of RRT* called Adapted RRT* to overcome the above limitations. The proposed approach found asymptotically optimal trajectories for a car-like robot with a 5D state space and an aerial vehicle with a 10D state space.

Recently, Lee et al. [44] proposed Spline-based RRT* (SRRT*) for non-holonomic path planning of fixed-wing Unmanned Aerial Vehicles (UAVs) in three-dimensional environment. The proposed algorithm expands the tree by using a cubic Bézier spline curve. Use of Bézier spline parameterization in SRRT* as local planner replaced time and input discretization. Moreover, it performed dynamic feasibility and geometric collision checks as part of the tree extension. This phenomenon enabled SRRT* to produce smooth and cost-optimal path in 3D simulations using constraint model of UAV.

Alejo et al. [3] presented RRT*i for efficient motion planning of UAVs. They compared the proposed approach with genetic algorithm, RRT and RRT*. Simulations results and real world experiments using UAV proved that RRT*i computes more predictable and smoother trajectories as compared to aforementioned techniques. Initially, RRT*i also works similar to RRT*. However, as soon as an initial trajectory is found, it adapts local sampling with Gaussian distribution and node rejection technique. Using this sampling technique it refines tree in vicinity of initial path solution, which leads to rapid convergence with less jagged path segments.

Recently, Csorvasi et al. [43] have proposed RTR+C*CS which uses a global planner and a local planner for car-like robots. First, global planner called RTR (Rotate-Translate-Rotate) generates path. Then local planner called C*CS makes it feasible for car-like robots using circular arcs and straight segments. However, its local planner requires large number of iterations to find less sharp turns.

Moon et al. [39] presented a kinodynamic variant of RRT* called Dual-Tree RRT (DT-RRT). DT-RRT manages two trees called state tree and workspace tree. At first, workspace tree explores targeted environment without considering any physical constraints. Then, state tree generates trajectories from workspace tree nodes using kinematic and dynamic constraints. It also offers reconnect-tree scheme in contrast to rewire operation of RRT* [7]. The reconnect-tree scheme [39] maintains child nodes with reduced computational cost, in order to guarantee asymptotic optimality. The dual tree scheme of DT-RRT [39] approach showed high success rate for node extension because it reduced the node rejection chances by applying kinodynamic conditions. Moreover, this approach is compatible with advanced near-neighbor search schemes, for example *k-d* trees with reduced space dimensionality. The proposed approach showed better computation time in simulation results using two wheeled mobile robot for high speed navigation. However, this scheme is only useful for known or partially known environment.

Another variant of RRT* proposed by Lan and Cairano [2] was experimented under Mitsubishi Electric Research Laboratories for autonomous driving vehicle. The proposed approach used two-stage sampling strategy similar to SRRT [40] and a weighted cost function tailored to drive semi-autonomous vehicle. The proposed approach could also manage driving lanes beside collision avoidance. A local re-planning procedure enables algorithm to react with dynamic obstacles. Further, path pruning and G^2 continuous curvature smoothing techniques are applied as post processing.

J. Suh et al. [65] presented an offline Cost Aware RRT* (CARRT*) for energy efficient optimal path planning in high dimensional space for humanoid robot. The approach used two trees to address the dense sampling issue in RRT*. First tree is a standard RRT* tree to determine nearest node for newly sampled random point whereas second tree contains first tree to extend additional long branches. They also proposed a cross entropy function based on CE path planning [66] and a cost function based on mechanical work (MW) [67] to measure energy consumption along a path for humanoid robot. Limitation of large memory requirements is also addressed by another variant called Potential Function RRT* (PRRT*) presented by Qureshi, and Ayaz [48]. PRRT* (PRRT*) is the extension of two other variants PGD-RRT* and APGD-RRT* [68]. As compared to these two approaches, PRRT* efficiently integrates artificial potential field with RRT* for guided exploration of search space with less memory needs and improved convergence.

Devaurs et al. [46] proposed Transition based RRT* (T-RRT*) which focuses on optimal path planning for continuous configuration-cost space. Their approach integrated transition test based functions used in T-RRT [67] to address the issue of path quality with respect to a given criterion.

Recently, Choudhury et al. [47] have proposed RABIT* to address the problem of planning in high dimensions. The proposed approach focuses homotopy classes which are difficult to sample for example narrow passages. RABIT* was extended by an informed global technique called BIT* [69] by using a local optimization module to improve an initial

sub-optimal path towards a local optimum. Thus, the proposed approach preserved almost-sure global optimal convergence.

IV. STATE OF THE ART (2011-2016)

The most relevant papers reviewed in this article, along with the research contributions and limitations are summarized below in Table 1. Different attributes of the state of the art approaches are also listed in Table 2.

TABLE I. STATE OF THE ART (2011-2016)

Sr#	Author, year	Approach	Research Contributions	Limitations / Future Recommendations
1.	Karaman and Frazzoli [7], 2011	RRT*	• Proved asymptotically optimal property for RRT*. • Introduced new key features of near neighbor search and rewiring operations. • Visibly refined path quality than original RRT.	• New features had an efficiency trade-off. Insertion of good candidate node with best parent selection improved tree cost but on the other hand it also slowed down convergence rate of RRT*. • Jagged, suboptimal paths and slow convergence • Large memory requirements.
2.	Karaman et al. [4], 2011	Anytime RRT*	• It introduced two key features called committed trajectories and branch-and-bound adaptation. • It improved trajectory and computational efficiency by gradually removing nodes from tree which are unable to improve current solution path.	• Jagged and suboptimal paths. • Could overestimate and may cause unnecessary node removal during initial expansion of tree when it is not mature.
3.	Akgun and Stilman [58], 2011	B-RRT*	• Improved convergence speed and path refinement using sample rejection with an admissible heuristic.	• Attempt to connect both trees in each iteration incurred computational overhead. • Jagged and suboptimal paths. • Large memory requirements.
4.	Adiyatov and Varol [33], 2013	RRT*FN	• Memory efficient version of RRT* by forcing a fixed number of nodes in the tree.	• Jagged, suboptimal path. • Rate of convergence to optimal path is lower than base RRT*. • Worked only for static known environment.
5.	Nasir et al. [35], 2013	RRT*-Smart	• Two new features intelligent sampling and path optimization were introduced. • It accelerated the convergence rate with improved efficiency with respect to both time and cost	• It is dependent upon a heuristic called Biasing Ratio which has a trade-off between convergence rate and exploration of space. • Heuristic used by this approach are not automated and require programmer dependent value for different environments.
6.	Jordan and Perez [36], 2013	Optimal B-RRT*	• Introduced multiple heuristics, based on different conditions to increase convergence rate of bidirectional RRT*.	• Use of multiple heuristics caused computational overload. • Biased heuristics interfere with the algorithm characteristics (such as exploration, node rejection, cost function) and limited its application also.
7.	O. Arslan and P. Tsiotras [50], 2013	RRT# (RRT "sharp")	• Introduced a global replanning scheme to maintain promising nodes of tree to make fast convergence towards optimal path with low cost.	• Real time experiments and applications would be beneficial to investigate and improve the efficiency of approach.
8.	Webb and Berg [49, 64], 2012, 2013	Adapted RRT*	• Presented kinodynamic extension of RRT* with ensured asymptotic optimality with controllable linear dynamics, in multi dimension state space.	• Post processing steps for path smoothness and real world quadrotors experiments are planned for future work. However, such experiment would require controller stabilization for final trajectory.
9.	Lee et al. [44], 2014	SRRT*	• Proposed spline based RRT* based on a cubic Bézier curve for fixed-wing UAVs. • Presented geometric collision and dynamic feasibility function checking constraints during tree expansion. • Produced feasible smooth and cost-optimal path in 3D simulations.	• Online planner with real time application would be beneficial to further investigate and improve the efficiency of approach.
10.	Gammell et al. [34], 2014	Informed RRT*	• Proposed direct sampling technique based on ellipsoidal informed subset, which showed improved convergence than RRT*.	• Heuristic used to shrink planning problem is highly dependent upon initial solution cost which makes it effective only under certain conditions.
11.	Qureshi and Ayaz [37], 2015	IB-RRT*	• Introduced intelligent sample insertion heuristic with minimal memory requirements, which improved the path quality as compared to RRT* and B-RRT*.	• Its application needs further investigation for online planning.
12.	Moon and Chung [39], 2015	DT-RRT	• Addressed kinodynamic planning for high speed mobile robot and produced practically feasible trajectories. • Instead of using rewire operation, it introduced reconnect-tree scheme to maintain child nodes with reduced computational cost.	• Approach need advancement for dynamic and unknown scenarios as its reconnect-tree scheme is only beneficial in known environments.

			• Reduced the node rejection chances by using kinodynamic conditions.	
13.	Alejo et al. [3], 2015	RRT*i	• Predictable and practical trajectory generation for UAVs. • Smoother trajectories as compared to RRT and RRT* with improved path quality. • Introduced new local sampling technique.	• Could produce unexpected collisions in multi-UAV applications with increased uncertainty.
14.	Csorvasi et al. [43], 2015	RTR+CS*	• Used a local planner to generate feasible path for car-like robots using circular arcs and straight segments.	• Generated paths are not curvature continuous and natural. • It is still under experimental process to improve computational performance.
15.	Lan and Di Cairano [2], 2015	Mitsubishi RRT*	• Addressed the problem of G^2 continuous curvature smooth path for autonomous driving vehicle with the capability of lane management on roads. • Introduced a local replanning procedure to safely avoid and re-plan due to dynamic obstacles.	• In context of autonomous vehicle driving, uncertainty of dynamic environment is not managed.
16.	J. Suh et al. [65] , 2015	CARRT*	• Addressed the problem of large number of samples by using two trees and cross entropy in RRT*. • Produced energy efficient path in high dimensional space.	• It is limited to address the planning problems for humanoid robots only.
17.	Qureshi and Ayaz [48], 2016	PRRT*	• Addressed the problem of high memory consumption and slow convergence by incorporating artificial potential field characteristic in RRT*.	• Considering optimal results with fast convergence than RRT*, approach could be employed for online planning in future.
18.	Yi et al. [45], 2016	HARRT*	• Presented a human-robot interactive planner using RRT* and homotopy algorithm.	• Trade-off between computational efficiency and path quality. • Natural language processing or graphical user interface could be adapted for compatibility, usability and workload of human and robot interactions.
19.	Devaurs et al. [46], 2016	T-RRT*	• Integrated transition tests with RRT* for efficient extension of tree in a cost space.	• Performance analysis with RRT* in different problems would be of interest to show its beneficial problem class.
20.	Choudhury et al. [47], 2016	RABIT*	• Used an informed global technique BIT* with RRT* to find optimal path for narrow passages in high dimensions.	• Local optimizer suggested by approach needs further investigation for optimization of path.

TABLE II. ATTRIBUTE SUMMARY OF THE STATE OF THE ART (2011-2016)

Approaches	Constraints	Planning Mode	Kinematic Model	Sampling Strategy	Metric
1. RRT* [7]	Holonomic	Offline	Point	Uniform	Euclidean
2. Anytime RRT* [4]	Non-holonomic	Online	Dubin Car	Uniform	Euclidean + Velocity
3. B-RRT* [58]	Holonomic	Offline	Rigid Body	Local bias	Goal biased
4. RRT*FN [33]	Holonomic	Offline	Robotic Arm	Uniform	Cumulative Euclidean
5. RRT*-Smart [35]	Holonomic	Offline	Point	Intelligent	Euclidean
6. Optimal B-RRT* [36]	Holonomic	Offline	Point	Uniform	Euclidean
7. RRT# [50]	Holonomic	Offline	Point	Uniform	Euclidean
8. Adapted RRT* [64], [49]	Non-holonomic	Offline	Car-like and UAV	Uniform	A* Heuristic
9. SRRT* [44]	Non-holonomic	Offline	UAV	Uniform	Geometric + dynamic constraint
10. Informed RRT* [34]	Holonomic	Offline	Point	Direct Sampling	Euclidean
11. IB-RRT* [37]	Holonomic	Offline	Point	Intelligent	Greedy + Euclidean
12. DT-RRT [39]	Non-holonomic	Offline	Car-like	Hybrid	Angular + Euclidean
13. RRT*i [3]	Non-holonomic	Online	UAV	Local Sampling	A* Heuristic
14. RTR+CS* [43]	Non-holonomic	Offline	Car-like	Uniform + Local Planning	Angular + Euclidean
15. Mitsubishi RRT* [2]	Non-holonomic	Online	Autonomous Car	Two-stage sampling	Weighted Euclidean
16. CARRT* [65]	Non-holonomic	Online	Humanoid	Uniform	MW Energy Cost
17. PRRT* [48]	Non-holonomic	Offline	P3-DX	Uniform	Euclidean

18. HARRT* [45]	Holonomic	Offline	Point	Uniform	Homotopy check
19. T-RRT* [46]	Non-holonomic	Offline	Quadrotor	Uniform	Transition test
20. RABIT* [47]	Non-holonomic	Offline	Autonomous helicopter	Uniform	A* Heuristic

V. Limitations and Prospect Challenges

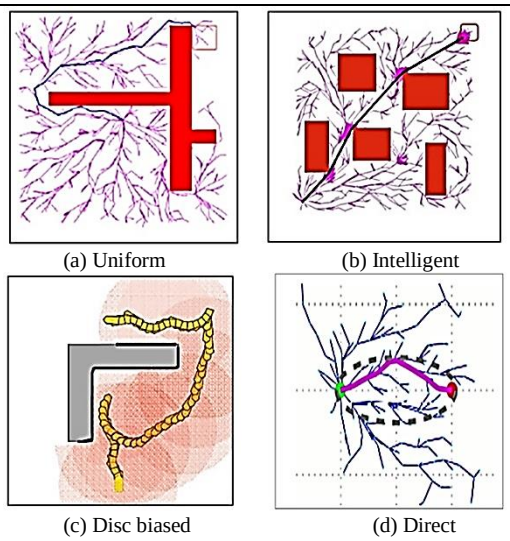
The existing state of the art requires improvement particularly in terms of accuracy, efficiency, robustness, and path optimization. Optimal path planning is a challenging problem and for online planning applications convergence to optimal path is even more important. This section describes the limitations addressed by variants of RRT* in recent years and also highlights the incessant future challenges.

A. Slow Convergence and Large Memory Requirements

RRT* requires large number of iterations and samples to avoid local minima consequently increasing memory requirements [48, 65]. Pure exploration also expands search space exponentially to find global optimum [47]. RRT* was proposed initially using uniform sampling strategy which was unable to effectively capture the connectivity of environment [7]. Further it also expands tree in the areas of configuration space that are far away from the final solution. Hence, a large number of nodes in tree are not good enough to contribute in the optimal path. These large number of nodes increase tree density by adding non-promising branches in tree. This phenomenon increases computational time and reduces convergence rate. Hence, slow convergence is also linked with search space exploration criteria and sampling strategy used by planner.

It is evident from discussion in Section III that dense sampling, large memory requirements and slow convergence are proven issues in RRT*. Recent RRT* based planners have addressed these issues by exploiting search space using different sampling strategies such as *direct sampling* [34], *goal biased sampling* [11, 56, 58], *intelligent sampling* [28, 35, 70], *two-stage sampling* [40], and *disc based sampling* [39, 42]. Few of them are shown in Fig. 6. The sampling strategies shown in Fig. 6 use different space explorations criteria. They try to limit search space using different heuristics to grow only promising branches and nodes in the tree. Another strategy reported in [48] performs guided exploration using artificial potential field to solve these issues. Moreover, different node deletion [33] or node rejection techniques [37] are also used to limit the tree cost by maintaining promising nodes according to a defined criterion.

However, these solutions especially guided exploration based solutions also require a careful balance of exploration and exploitation in search space. Moreover, sampling strategies based solutions use manual heuristics, which need specific tuning according to application or environment type for better performance. Therefore, all the strategies discussed above need further improvement regarding robustness and automation of heuristics parameters, specifically in the context of online planning.



(a) Uniform (b) Intelligent

(c) Disc biased (d) Direct

Fig. 6. Different sampling strategies to overcome slow convergence [35, 42]

B. Dealing with Narrow Passages

Conventional uniform sampling of RRT* reduces the probability of nodes selection from narrow passages. Very few approaches are reported to specifically address the problem of optimal path in narrow passages. These approaches either use heuristics dependent upon initial solution such as Informed RRT* [34] or they are under theoretical assumptions such as HARRT* [45]. There is a need to investigate the potential of these approaches to achieve optimization and reliability in narrow passages using real world experiments. Moreover, in context of kinodynamic constraints, problem of narrow passages is still an open research issue.

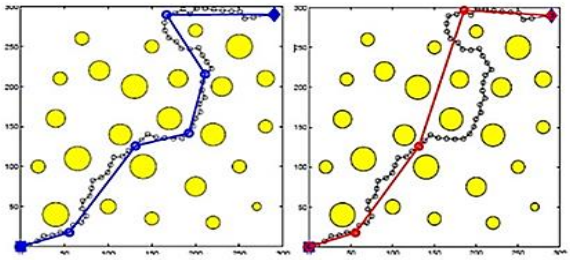
C. Efficiency of Nearest Neighbor Search

Computational complexity of near neighbor search in each iteration also grows as tree expands exponentially. Therefore, it is considered a bottleneck for efficiency and convergence. Adiyatov and Varol [33] maintained the efficiency by fixing maximum possible nodes in tree. Other strategies that most of the researchers have adopted are to use smarter search techniques such as Box approach [71] or smarter data structure such as *k-d* tree and quad trees [71]. Yershova and LaValle [72] proposed *k-d* trees based near neighbor search algorithm for Euclidean spaces. However, alternative techniques to promote least cost connections in tree or faster search algorithms could be helpful to further improve the efficiency.

D. Post Processing Requirements

As RRT* based approaches generate sub-optimal path therefore, post processing techniques are adopted to further optimize the path. Two post processing techniques usually adopted for path refinement are pruning and smoothing. Path pruning reduces the path length by removing redundant nodes [56]. Two types of pruning, local pruning and global pruning

are shown in Fig. 7 (a) and Fig. 7 (b) respectively. Local pruning is based on Line of Sight (LOS) principle whereas, global pruning removes the myopic behavior of local pruning by pruning the nodes of entire path [56].



(a) Local pruning (b) Global pruning

Fig. 7. Post processing schemes [40]

Even after pruning, generated linear piece-wise path is not feasible for UAVs and car-like robots. Curvature discontinuities in path make controller unstable and overshoot energy [41]. This phenomenon can cause mechanical aging, localization errors and high energy requirements. The situation becomes more complicated when planning application involves heavy machinery such as industrial, defense or agricultural robots.

To resolve the issues stated above, different levels of smoothing such as path smoothing, continuous smoothing, and continuous curvature smoothing [41] have been applied as post processing steps. Approaches used for this purpose are classified as graphical methods (lines, arcs, circles, and clothoids) and functional methods (Bézier, B-spline and polynomial interpolation). Recent RRT* variants have used Bézier and B-spline to meet the challenges of kinodynamic planning for non-holonomic vehicles effectively in [40], see Fig. 7(c).

However, these approaches also have the limitation of maintaining continuity using suitable degree of curve. Clamped B-spline is more robust for path smoothing than Bézier and B-spline due to its ability of maintaining continuity and order for dynamic re-planning [41]. Recently Elbanhawi et al. [41] have proposed a C^2 continuous path smoothing approach using clamped B-spline for continuous steering of car-like robots. Their approach mimics human steering with high accuracy using a threshold angle and segment insertion technique. Their smoothing approach could be applied with recent RRT* variants for improved performance.

However, dealing with non-holonomic and kinodynamic constraints after trajectory generation as post processing step increases the complexity of the planning and search space. Moreover, it is also computationally expensive due to frequent updates considering real world applications and online planners.

E. Dealing with Kinodynamic Complexities

Non-holonomic constraints require addition of robot's orientation to state vector. Thus, increased dimension increases complexity of configuration space exploration. Moreover, they involve solving differential equations and has more complicated state transition equation. Thus, planning for non-holonomic robots with kinodynamic constraints is more difficult and challenging than holonomic robots [1, 73]. Therefore, RRT* application with non-holonomic motion requires more iterations to converge than a holonomic version [57]. Usually path planners generate linear piece-wise path ignoring all kinodynamic constraints and path smoothing is applied later as post processing step as discussed above in section 4.4. Recent state of the art [41] is inclined to apply kinematic constraint model while searching the path to avoid complexities of post processing [8, 60].

RRT* based approaches have a steering function that connects configurations using a straight line from $z_{nearest}$ to z_{rand} to generate z_{new}. However, such a steering function is not feasible for non-holonomic robots. To resolve these issue kinematic model of non-holonomic robot is used, which involves numerical integration [40]. However, there is always a trade-off between computational efficiency and accuracy when using numerical integration. Recently, Lee et al. used spline in SRRT* [44] to resolve these issues by encoding non-holonomic constraints in spline generation during path searching process of planner. This process also reduced the kinodynamic planning to lower dimensional space.

RRT* defines a metric in configuration space to identify the nearest neighbors of a given configuration. Since robots have to meet the challenges of un-certain dynamic environment, actuator constraints, localization errors and computation limitations. There is no silver bullet managing all these factors and it is difficult to distinctively define such metric. In context of non-holonomic planning, simple Euclidean distance metric is not proficient to capture cost of node in configuration space. Tree nodes need to qualify feasibility test for non-holonomic constraints, which is quite a challenging task. Lee et al. [44] have used a spline based metric in SRRT* which performs dynamic feasibility and geometric collision checks. However, the proposed approach takes several minutes in simulation and in real-time it is even more delayed [73].

Poor selection of metric for non-holonomic planning could also limit robot motion in narrow passage and slow down the planner [62]. Therefore, metric should be a true representative of the effort, or time-to-go between two configurations, otherwise highly sub-optimal paths are produced [8]. Future work in metric design solving kinodynamic problem with real time computational efficiency is an important element to improve performance of RRT*. Moreover, path smoothness could be further improved using other Computer Aided and Graphic Design (CAGD) tools such as NURBS or spiral segments. Defining a suitable weight exploitation criteria for NURBS can control local change in path segments while performing path smoothness. Cubic Bézier spiral segments reduce the number of curvature extrema in path [74]. Therefore, exploiting curvature continuous spiral transition curves could be useful to produce smooth path for high speed moving objects [75] such as UAVs.

VI. Conclusion and Future Directions

Research over the past decade has revealed that traditional path planning methods are not feasible for non-holonomic,

cluttered and high dimensional problems. RRT* have proved its worth for dealing with such complex problems. This paper presents review of major contributions in optimal path planning using RRT* planning algorithm and its extended variants in recent six years. RRT* based approaches have revolutionized the state of the art in path planning. These recent variants have mostly addressed the issues of sub-optimal paths, slow convergence and high memory requirements. However, generalization and reliability in context of online planners and non-holonomic constraints are still open research issues. Though RRT* based approaches addressing kinodynamic and non-holonomic constraints are also in progress, however kinodynamic planning still confronts issues of high dimension, narrow passages and trade-off between accuracy and computational efficiency.

In recent planners, use of spline with RRT* has opened new horizon of research for path planning of non-holonomic robots. Curvature continuous path for high speed vehicles while considering non-holonomic constraints, uncertainty of dynamic environment, and preserving computational efficiency is another thriving area of future research. Use of other CAGD tools such as clamped B-spline, NURBS and spiral segments with RRT* variants for path planning of non-holonomic robots would be an interesting research endeavor in future.

REFERENCES

[1] S. M. Lavalle, Planning algorithms: Cambridge University Press, 2006.

[2] X. Lan, and S. Di Cairano, "Continuous curvature path planning for autonomous vehicle maneuvers using RRT*", presented at the European Control Conference (ECC), 2015.

[3] Alejo, J. A. Cobano, G. Heredia, J. R. Martínez-De Dios, and A. Ollero, "Efficient trajectory planning for wsn data collection with multiple UAVs", in Cooperative robots and sensor networks. vol. 604, ed: Springer International Publishing, 2015, pp. 53-75.

[4] S. Karaman, M. Walter, A. Perez, E. Frazzoli, and S. Teller, "Anytime motion planning using the RRT*", presented at the IEEE International Conference on Robotics and Automation (ICRA) 2011.

[5] *Lau, and H. H. T. Liu, "Real-time path planning algorithm for autonomous border patrol: Design, simulation, and experimentation", J Intell Robot Syst, vol. 75, pp. 517-539, 2013.*

[6] X. Kong, X. Duan, and Y. Wang, "An integrated system for planning, navigation and robotic assistance for mandible reconstruction surgery", INTEL SERV ROBOT, vol. 9, pp. 113-121, 2015.

[7] S. Karaman, and E. Frazzoli, "Sampling-based algorithms for optimal motion planning", Int J Rob Res, vol. 30, pp. 846-894, 2011.

[8] M. Elbanhawi, and M. Simic, "Sampling-based robot motion planning: A review survey", IEEE Access, vol. 2, pp. 56-77, 2014.

[9] C. Goerzen, Z. Kong, and B. Mettler, "A survey of motion planning algorithms from the perspective of autonomous UAV guidance", J INTELL ROBOT SYST, vol. 57, pp. 65-100, 2009.

[10] M. Nosrati, R. Karimi, and H. A. Hasanvand, "Investigation of the * (star) search algorithms characteristics methods and approaches", World appl program, vol. 2, pp. 251-256, April 2012.

[11] J. J. Kuffner, and S. M. Lavalle, "RRT-connect: An efficient approach to single-query path planning", in Proceedings of IEEE International Conference on Robotics and Automation (ICRA), 2000, pp. 1-7.

[12] W. Dijkstra, "A note on two problems in connexion with graphs", NUMER MATH, vol. 1, pp. 269-271, 1959.

[13] Zelinsky, R. A. Jarvis, J. C. Byrne, and S. Yuta, "Planning paths of complete coverage of an unstructured environment by a mobile robot", in Proceedings of International Conference on Advanced Robotics (ICAR), 1993.

[14] W. Zhan, W. Wang, N. Chen, and C. Wang, "Efficient UAV path planning with multiconstraints in a 3d large battlefield environment", MATH PROBL ENG, vol. 2014, pp. 1-12, 2014.

[15] Stentz, "Optimal and efficient path planning for partially-known environments", presented at the Proceedings of the International Conference on Robotics and Automation, 1994.

[16] Nash, S. Koenig, and M. Likhachev, "Incremental phi*: Incremental any-angle path planning on grids", in International Joint Conference on Aritificial Intelligence, 2009, pp. 1824–1830.

[17] K. Sameshima, K. Nakano, T. Funato, and S. Hosokawa, "StRRT-based path planning with pso-tuned parameters for robocup soccer", AROB, vol. 19, pp. 388-393, 2014.

[18] P. Garcia, O. Montiel, O. Castillo, R. Sepúlveda, and P. Melin, "Path planning for autonomous mobile robot navigation with ant colony optimization and fuzzy cost function evaluation", APPL SOFT COMPUT, vol. 9, pp. 1102-1110, 2009.

[19] J. Kennedy, and R. C. Eberhart, "Particle swarm optimization", in IEEE International Conference on Neural Networks (ICNN), 1995.

[20] M. Dorigo, Ant colony optimization: Cambridge University, Massachusetts: MIT Press, 1992.

[21] T. Arora, Y. Gigras, and V. Arora, "Robotic path planning using genetic algorithm in dynamic environment", INT J COMPUT APPL, vol. 89, pp. 8-12, 2014.

[22] J. H. Liang, and C. H. Lee, "Efficient collision-free path-planning of multiple mobile robots system using efficient artificial bee colony algorithm", ADV ENG SOFTW, vol. 79, pp. 47-56, 2015.

[23] M. A. Hossain, and I. Ferdous, "Autonomous robot path planning in dynamic environment using a new optimization technique inspired by bacterial foraging technique", ROBOT AUTON SYST, vol. 64, pp. 137-141, 2015.

[24] D. Q. Zhu, W. C. Li, M. Z. Yan, and S. X. Yang, "The path planning of auv based on d-s information fusion map building and bio-inspired neural network in unknown dynamic environment", International Journal of Advanced Robotic Systems, vol. 11, March 6 2014.

[25] X. Wang, Z. Hou, F. Lv, M. Tan, and Y. Wang, "Mobile robots′ modular navigation controller using spiking neural networks", Neurocomputing, vol. 134, pp. 230-238, 2014.

[26] Fister, I. Fister, X.-S. Yang, and J. Brest, "A comprehensive review of firefly algorithms", Swarm and Evolutionary Computation, vol. 13, pp. 34-46, 2013.

[27] Kenedy, Eberhart, and Shi, Swarm intelligence: Moragan Kaufmann Division of Academic Press, 2001.

[28] *Lin, and C. Yang, "2d-span resampling of bi-RRT in dynamic path planning", Int J Autom Smart Technol, vol. 4, pp. 39-48, 2015.*

[29] I. Tsianos, I. A. Sucan, and L. E. Kavraki, "Sampling-based robot motion planning: Towards realistic applications", Comput Sci Rev, vol. 1, pp. 2-11, 2007.

[30] E. Kavraki, P. Svestka, J. C. Latombe, and M. Overmars, "Probabilistic roadmaps for path planning in high dimensional configuration spaces ", IEEE Transactions on Robotics and Automation, vol. 12, pp. 566-580, 1996.

[31] S. M. Lavalle, "Rapidly-exploring random trees: A new tool for path planning", 1998.

[32] Snchez, L. Zapata, J. Abraham, and A. B., "Motion planning for car-like robots using lazy probabilistic roadmap method", in Proceedings of the Second Mexican International Conference on Artificial Intelligence: Advances in Artificial Intelligence, 2002, pp. 1-10.

[33] O. Adiyatov, and H. A. Varol, "Rapidly-exploring random tree based memory efficient motion planning", presented at the IEEE International Conference of Mechatronics and Automation (ICMA), 2013.

[34] J. D. Gammell, S. S. Srinivasa, and T. D. Barfoot, "Informed RRT*: Optimal sampling-based path planning focused via direct sampling of an admissible ellipsoidal heuristic", in IEEE RSJ International Conference on Intelligent Robots and Systems (IROS), Chicago, 2014, pp. 2997-3004.

[35] J. Nasir et al., "RRT*-smart: A rapid convergence implementation of RRT*", International Journal of Advanced Robotic Systems, vol. 10, pp. 1-12, 2013.

[36] Jordan, and A. Perez, "Optimal bidirectional rapidly-exploring random trees", Computer Science and Artificial Intelligence Laboratory, Massachusetts Institute of Technology, Cambridge, MA, Tech. Rep. MIT-CSAIL-TR-2013-021, August 2013.

[37] H. Qureshi, and Y. Ayaz, "Intelligent bidirectional rapidly-exploring random trees for optimal motion planning in complex cluttered environments", ROBOT AUTON SYST, vol. 68, pp. 1-11, 2015.

[38] K. Yang, S. K. Gan, and S. Sukkarieh, "A gaussian process-based RRT planner for the exploration of an unknown and cluttered environment with a UAV", ADV ROBOTICS, vol. 27, pp. 431-443, 2013.

[39] Moon, and W. Chung, "Kinodynamic planner dual-tree RRT (dt-RRT) for two-wheeled mobile robots using the rapidly exploring random tree", IEEE T IND ELECTRON, vol. 62, February 2015.

[40] K. Yang et al., "Spline-based RRT path planner for non-holonomic robots", J Intell Robot Syst, vol. 73 pp. 763-782, 2014.

[41] Elbanhawi, M. Simic, and R. Jazar, "Continuous path smoothing for car-like robots using b-spline curves", J Intell Robot Syst, pp. 1-34, 2015.

[42] Masehian, and H. Kakahaji, "Nrr: A nonholonomic random replanner for navigation of car-like robots in unknown environments", Robotica, vol. 32, pp. 1101-1123, 2014.

[43] Csorvasi, A. Nagy, and D. Kiss, "Rtr+c*cs: An effective geometric planner for car-like robots", presented at the 16th International Carpathian Control Conference (ICCC), 2015.

[44] Lee, H. Song, and D. H. Shim, "Optimal path planning based on spline-RRT* for fixed-wing UAVs operating in three-dimensional environments", presented at the 14th International Conference on Control, Automation and Systems (ICCAS 2014), Korea, 2014.

[45] Yi, M. A. Goodrich, and K. D. Seppi, "Homotopy-aware RRT* : Toward human-robot topological path-planning", presented at the The 11th ACM/IEEE International Conference on Human Robot Interaction (HRI), Christchurch, New Zealand, 2016.

[46] Devaurs, T. Simeon, and J. Cortes, "Optimal path planning in complex cost spaces with sampling-based algorithms", IEEE T AUTOM SCI ENG, vol. 13, pp. 415-424, Apr 2016.

[47] S. Choudhury, J. D. Gammell, T. D. Barfoot, S. S. Srinivasa, and S. Scherer, "Regionally accelerated batch informed trees (rabit*): A framework to integrate local information into optimal path planning", presented at the IEEE International Conference on Robotics and Automation (ICRA), Stockholm, Sweden, 2016.

[48] H. Qureshi, and Y. Ayaz, "Potential functions based sampling heuristic for optimal path planning", AUTON ROBOT, vol. 40, pp. 1079-1093, 2016.

[49] D. J. Webb, and J. V. D. Berg, "Kinodynamic RRT*: Asymptotically optimal motion planning for robots with linear dynamics", presented at the IEEE International Conference on Robotics and Automation (ICRA), Karlsruhe, Germany, 2013

[50] Arslan, and P. Tsiotras, "Use of relaxation methods in sampling-based algorithms for optimal path planning", presented at the IEEE International Conference on Robotics and Automation (ICRA), Karlsruhe, Germany, 2013.

[51] Y. K. Hwang, "Gross motion planning-a survey", ACM COMPUT SURV, vol. 24, pp. 219-291, 1992.

[52] Noreen, A. Khan, and Z. Habib, "A comparison of RRT, RRT* and RRT*-smart path planning algorithms", IJCSNS, vol. 16, pp. 20-27, 2016.

[53] W. Loeve, "Finding time-optimal trajectories for the resonating arm using the RRT* algorithm", Master of Science, Faculty Mechanical, Maritime and Materials Engineering, Delft University of Technology, Delft, 2012.

[54] *S. Yoon, and T. H. Park, "Motion planning of autonomous mobile robots by iterative dynamic programming", INTEL SERV ROBOT, vol. 8, pp. 165-174, 2015.*

[55] D. Ferguson, and A. Stentz, "Anytime, dynamic planning in high-dimensional search spaces", presented at the International Conference on Robotics and Automation (ICRA), Italy, 2007.

[56] Yang, "Anytime synchronized-biased-greedy rapidly-exploring random tree path planning in two dimensional complex environments", INT J CONTROL AUTOM, vol. 9, pp. 750-758, 2011.

[57] R. Nguyen. (2010). Project adv - autonomous driving vehicle. Available: http://webpages.uncc.edu/nhnguye1/ADV.html

[58] B. Akgun, and M. Stilman, "Sampling heuristics for optimal motion planning in high dimension", presented at the International Conference on Intelligent Robots and Systems, San Francisco, CA, USA, 2011.

[59] N. Ahmidi, G. D. Hager, L. Ishii, G. L. Gallia, and M. Ishii, "Robotic path planning for surgeon skill evaluation in minimally-invasive sinus surgery", presented at the 15th International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI), 2012.

[60] B. Donald, P. Xavier, J. Canny, and J. Reif, "Kinodynamic motion planning", J Acm, vol. 40, pp. 1048-1066, Nov 1993.

[61] T. Fraicharda, and A. Scheuerb, "From reeds and shepp's to continuous-curvature paths", IEEE T ROBOTIC AUTOM, vol. 20, 2004.

[62] S. Karaman, and E. Frazzoli, "Optimal kinodynamic motion planning using incremental sampling-based methods", presented at the 49th International Conference on Decision and Control (CDC), 2010.

[63] Perez, R. Platt, G. Konidaris, L. Kaelbling, and T. Lozano-Perez, "Lqr-RRT*: Optimal sampling-based motion planning with automatically derived extension heuristics", in IEEE International Conference on Robotics and Automation, 2012, pp. 2537-2542.

[64] D. J. Webb, and J. V. D. Berg, "Kinodynamic RRT*: Optimal motion planning for systems with linear differential constraints", arXiv:1205.5088v1, 2012.

[65] Suh, J. Gong, and S. Oh, "Energy efficient high dimensional motion planning for humanoids using stochastic optimization", presented at the 15th International Conference on Humanoid Robots, Seoul, Korea, 2015.

[66] Kobilarov, "Cross-entropy randomized motion planning", in Proceedings of Robotics: Science and Systems, Los Angeles, CA, USA, 2011, pp. 1-8.

[67] Jaillet et al., "Sampling-based path planning on configuration-space costmaps", IEEE T ROBOT, vol. 26, pp. 635-646, 2010.

[68] H. Qureshi et al., "Potential guided directional-RRT* for accelerated motion planning in cluttered environments", in IEEE International Conference on Mechatronics and Automation, 2013, pp. 519-524.

[69] *D. Gammell, S. S. Srinivasa, and T. D. Barfoot, "Batch informed trees (bit*): Sampling-based optimal planning via the heuristically guided search of implicit random geometric graphs", presented at the IEEE International Conference on Robotics and Automation, 2015.*

[70] Islam, J. Nasir, U. Malik, Y. Ayaz, and O. Hasan, "RRT*-smart: Rapid convergence implementation of RRT* towards optimal solution", presented at the International Conference on Mechatronics and Automation (ICMA), Chengdu, 2012.

[71] Svenstrup, T. Bak, and H. J. Andersen, "Minimising computational complexity of the RRT algorithm a practical approach", presented at the IEEE International Conference on Robotics and Automation, Shanghai, China, 2011.

[72] Yershova, and S. M. Lavalle, "Improving motion planning algorithms by efficient nearest-neighbor searching", IEEE T ROBOTIC AUTOM, vol. 23, pp. 151-157, 2007.

[73] S. M. Lavalle, and J. J. Kuffner, "Randomized kinodynamic planning", Int J Rob Res, vol. 20, pp. 378-400, May 2001.

[74] Z. Habib, and M. Sakai, "Fairing arc spline and designing by using cubic Bézier spiral segments", MATH MODEL ANAL, vol. 17, pp. 141-160, 2012.

[75] Z. Habib, Spiral function and its applications in CAGD: VDM, 2010.

Performance Analysis of In-Network Caching in Content-Centric Advanced Metering Infrastructure

Nour El Houda Ben Youssef
SAGE-LATIS/ ENISO University of Sousse
ENSI University of La Manouba
WEVIOO
Tunisia

Yosra Barouni, Sofiane Khalfallah, Jaleleddine Ben Hadj Slama
SAGE-LATIS/ENISO University of Sousse Tunisia

Khaled Ben Driss
WEVIOO
Tunisia

***Abstract*—In-network caching is a key feature of content-centric networking. It is however a relatively costly mechanism with hardware requirements besides placement/replication strategies elaboration. As content-centric networking is proposed in the literature to manage smart grid (SG) communications, we aim, in this research work, to investigate the cost effectiveness of in-network caching in this context. We consider, in particular, the Advanced Metering Infrastructure (AMI) service that comes into prominence since its outputs are imperative inputs of most smart grid applications. In this research work, AMI communication topology and data traffic are characterized. Corresponding simulation environment is then built. Thereafter, various placement and replacement strategies are compared in a simulation study to be further able to propose a suitable cache placement and replacement combination for AMI in Smart Grid.**

Keywords—caching; placement; replacement; content-centric networking; Named Data Networking; Advanced Metering Infrastructure; Smart Grid

I. INTRODUCTION

As a heterogeneous, distributed and large scale system, the Smart Grid Communication System (SGCS) is the subject of many research works trying to propose a well-tailored communication layer that guarantees smart grid requirements. Among many solutions, CCN is proposed to offer real time data transmission with inherent security levels and competitive latency. This research work belongs to a series of articles aiming at investigating content-centric networking adequacy for smart grids. In[1], CCN performances were compared to the Internet Protocol stack while managing smart grid communications. A deeper qualitative and quantitative analysis is pursued in [2] studying CCN support for renewable energy resources integration into SGs. The aforementioned works allowed us to take a position in favor of CCN as an eligible communication solution for this system. In this article, the focus is granted to caching as one of the most important building blocks of CCN. In fact, besides host decoupling and content-based routing, in-network caching is considered as one of the most relevant CCN features. The main goal is to study the cost effectiveness and the performance of this mechanism in smart grids. Advanced Metering Infrastructure (AMI), a service responsible for the generation of a huge amount of data such as power consumption and electrical parameters data, is particularly considered. This smart grid application can particularly avail of in-network caching since many data flows disseminate the same content to many nodes like price signals or maintenance commands. Previous work [3] proposes CCN for enabling AMI service but the focus is mostly on proposing a naming scheme to elaborate CCN-AMI. A relatively poor performance assessment is presented as only bandwidth consumption metric is studied. To the best of our knowledge, no other research work dealt with content-centric in-network caching impact on AMI performance.

This article leads off by an overview of the advanced metering infrastructure detailing this service and its major stakeholders. Afterward, part II exposes related works by exploring research works investigating CCN use in smart grids. Then content-centric networking is presented with a particular focus on its in-network caching feature in part III. Next, simulation environment and scenarios description is presented in part IV. The last part of this article exposes and analyses the results obtained by the conducted simulations.

II. STATE OF THE ART: CCN IN SMART GRIDS

Content-centric networking (CCN) in smart grids is being in the center of an active research effort where researchers and industrials are collaborating to propose a content-centric communication layer for smart grids. The first endeavor to enable SGCS using CCN was in [4] where a content-centric overlay network is deployed to exchange smart grid data traffic. It, however, used geographic routing preventing it to be considered as a pure content-centric solution. Then, the authors of [5] studied Information-Centric Networking (ICN) performance in Real Time State Estimation (RTSE) of smart grids. This work has been expanded in [6] and an information-centric platform baptized C-DAX is proposed to RTSE in active power distribution grids. Other research works dealt with different smart grid subsystem, we noticed particularly the adoption of ICN in home area networks for residential energy management [7]. ICN is also investigated to enable smart city services [8]; in particular vehicle-to-grid communications has been studied. The aforementioned research works analyzed ICN performance in selected smart grid subsystems but, despite its importance, no particular attention is paid to caching policies assessment. Due to its deep impact on networking performances, caching mechanisms requires thorough analysis. The present research

This research work is elaborated under the umbrella of the PASRI-MOBIDOC (www.pasri.tn) project in Tunisia funded by the European Union and WEVIOO (www.wevioo.com).

work aims at filling this gap by probing into in-network caching impact on AMI communications.

III. Advanced Metering Infrastructure

A. Overview

Advanced metering infrastructure (AMI) is a smart grid potential application on the consumer side, responsible for metering services between electricity utility companies and their customers. It is an integrated system of smart meters, communication channels and meter data management systems. Many relevant smart grid subsystems like Advanced Distribution Infrastructure (ADI), Advanced Transmission Infrastructure (ATI) and Asset Management (AM) rely on the advanced metering infrastructure in order to properly achieve their functions. AMI has many benefits to both customers and power providers. On the consumer side, it allows him to be well informed regarding prices and services which provide him with the ability of managing its consumption patterns and costs. On the power utility side, AMI has an impact on two major services: the billing and Distribution/Transmission operations. Receiving time stamped power consumption information helps establishing an efficient billing system. In addition, customer information processing and mining may enable the utility offerings improvements. Moreover, accurate reporting of various electrical parameters is essential for the smart grid main purpose which is a reliable power transmission and distribution.

B. Actors, roles and interactions

Realizing the power grid modernization relies on several cooperating subsystems. Deploying an advanced metering infrastructure can be considered as one imperative step toward the smart grid since its outputs are the feeds of advanced distribution and transmission infrastructures [9]. Most relevant AMI functions are the following:

- Power data consumption readings,
- Electrical parameters reporting (phase, voltage, power factor),
- Time-based pricing and billing,

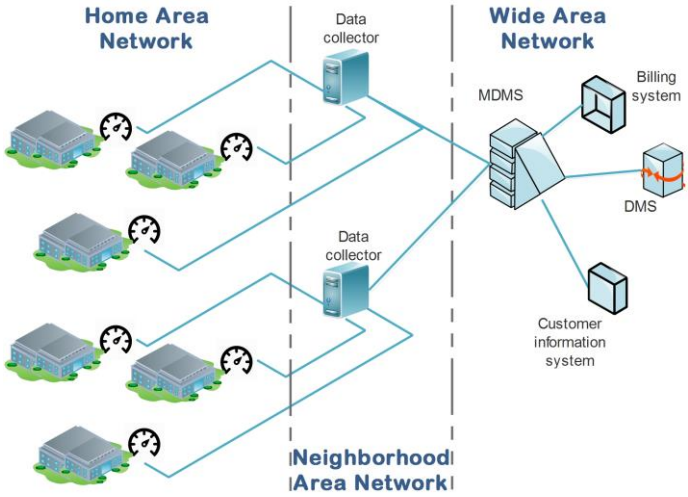


Fig. 1. Architecture of an AMI

Achieving the aforementioned functionalities involves four main actors (see Fig.1):

- **Smart meters**: a device deployed at the customer premises responsible for periodic power consumption data readings. It also contributes at transmission and distribution operations by reporting some electrical parameters like phase, voltage, etc. Smart meters have to meet few requirements; communication skill can be considered as the most imperative one as it allows the device to achieve its major role. Smart meters must provide users by an ergonomic display feature. It also should support remote control commands to be executed by distribution management systems in order to maintain a reliable power grid.
- **Electrical properties sensors**: Voltage sensors (VS) and phase measurement units (PMU) are electrical field devices responsible for reporting real time electrical properties used to monitor the power distribution system.
- **Communication layer**: Each smart meter deployed in the smart grid has IN/OUT data traffic to receive or send. It has, in fact, to periodically send data about power consumption or electrical parameters status to remote collecting and processing nodes. In addition, it receives operational commands and price signals from operation centers. A communication layer managing all these data flows is then an important AMI subsystem. The richness of communication technologies landscape allows many design choices. Wireless and wired technologies can be combined to build a well tailored communication layer considering bandwidth and latency requirements at each level of the grid.
- **Data management system**: It is an integrated system for managing the huge amount of data received from customers at the utility company premises. Data collectors are first deployed in neighborhoods to collect data from smart meters in Home Area networks (HANs). It includes then meter data management system (MDMS) that can be seen as a database responsible for gathering and storing metering data from data collectors in neighborhood area networks (NANs). This system benefits, nowadays, from big data technologies and data centers expansion. The data management system also includes a billing system and a customer information system and has several interfaces with distribution operators such as distribution management system. Indeed this latter needs power quality information sent by smart devices deployed throughout the grid in order to monitor the power distribution system and trigger operational commands if needed.

To achieve advanced metering functions, the aforementioned actors need to interact and exchange data. The most relevant data flows are power consumption information sent periodically by smart meters and collected in data collectors deployed in neighborhoods. Power consumers receive also real time power pricing allowing them to have an

active role regarding their consumption patterns. AMI data traffic description provided in Table 1 was elaborated after consulting various research articles surveying smart grid services [10][11].

IV. Caching in Content-Centric Networking

A. Content-centric networking in a nutshell

Till now, communication solutions proposed to manage computer networks are mostly based on addressing hosts and maintaining a bi-host connection used to exchange data. However, new communication patterns are content dissemination oriented which lead to the imperative need of a recast of existing networking paradigms. This fact has enriched the literature, since the early 2000s, by several research works [12][13][14] proposing the recast of the Internet and the adoption of new communication tenets. Although content-centric networking has first been elucidated in [15] to cater for the Internet recast, it has been explored by many researchers to grant required networking performances in diverse use cases such as social networking, vehicular networking, smart cities, etc. This attraction is due to many reasons. Building the whole communication process around the content unlike traditional paradigms where hosts are in the center of communications can be considered as the major attractive treat of CCN. Indeed we pay more attention on retrieving the desired content then knowing the host providing it. To do so content needs to be identified instead of addressing hosts this guaranteed by a naming strategy: the second pillar of CCN. A content-based routing protocol is also used to manage content travel from its producer to its consumer. In-network caching, an important building block of CCN, ardently seduces researchers as it promises enhanced networking performances.

B. In-network caching

CCN is characterized by in-network caching of data that improves network quality of services especially delivery latency. Satisfying data requests is not obligatory through locating the original data source but can also be done from multiple data stores. We find, in the literature, many caching policies that can be categorized in in-path caching or off-path caching. In the first one, only data replicas found along the path taken by the name resolution request are exploited. In contrast, the second one allows exploiting caches outside this path [17]. We can also differentiate caching policies based on router cooperation while placing content cache. Cooperative or non-cooperative strategies are mostly known [16].

In order to achieve in-network caching, placement and replacement algorithms need to be deployed. First of all, one needs to decide where to cache content; would it be in each node or in some selected ones. This decision leads us to the elaboration of a cache placement strategy like Leave Copy Everywhere (LCE) or Leave Copy Down (LCD), etc [18]. It is also common to place caches probabilistically or based on content proximity to its consumer [19]. A cache replacement strategy is also required since we wouldn't like to exceed the cache size. Least Recently Used (LRU), Least Frequently Used (LFU) and First In First Out (FIFO) are the most usual replacement policies in the literature.

V. Performance Analysis of Caching for Advanced Metering Infrastructure

A. Simulation tools

Being in the heart of an intensive research effort, content-centric networking is the subject of many research projects which lead us to many implementations of this paradigm. Among many, Named Data Networking (NDN) project (https://named-data.net/) is taking the lead by producing a large panel of deliverables such as core solution implementation, periodic technical reports, testbeds and simulation modules. In the present research work, we use NDNSIM, the NDN simulation module under NS-3, to simulate content- centric communications. Network simulator-3 (NS-3): https://www.nsnam.org/) is an event-based simulator widely used by the research community due to the richness of its libraries. We first built a networking topology according to figure 1 in order to reproduce AMI communication environment. Since no consensus has been made on the communication technologies to use while building the smart grid communication system, we have abstracted from the physical layer. In fact, the networking landscape offers many alternatives to enable smart grid communications; for instance home area networks (HANs) can exploit narrowband power line communication as wired technology or Bluetooth or Wi-Fi as wireless technology. The common aspect at this level of the grid is a low data rate. As for neighborhood area networks (NANs), cellular networks of broadband power line communications can be adopted. Finally the wide area network (WAN) can be achieved by optical fiber technology or fourth and fifth cellular network generations such as LTE [20]. Since assessing physical technologies performances in smart grids requires a dedicated research work, we excluded this task from our scope and therefore decided to use point to point links in NS-3 with well studied characteristics according to the requirements of each smart grid communication level.

TABLE I. AMI DATA FLOWS

Id	Source Actor	Exchanged Information	Destination Actor	Packet size (bytes)	Frequency
1	Smart meter	Power consumption data	Meter data mangement system (MDMS)	200	Every hour
2	Billing system	Real time prices	Smart meter	210	Every 15 minutes
3	Distribution management system (DMS)	Operationl commands	Smart meter	150	Triggered by events
4	Voltage sensor (smart meter)	Voltage, power factor	Distribution management system (DMS)	250	Every 5 minutes
5	Voltage sensor (transformer)	Voltage, power factor	Distribution management system (DMS)	250	Every 5 minutes
6	Phase measurment unit (transformer)	Voltage Phase	Distribution management system (DMS)	1536	Every 5 minutes

B. Simulation topology

To conduct simulations, the AMI topology has been built under NS-3 with the restriction of our scope to 5 Km radius area. Based on a prior research work[11] describing a smart grid communication infrastructure in Canada, it has been noticed that the communication infrastructure generally follows the electrical one which implies that the smallest branch of the adopted topology contains 10 smart meters according the north American norms [18]. In fact, within 5 km, 50 substations are deployed and 120 transformers are connected to each substation. Then, 10 homes will be served by each transformer leading to 10 smart meters wired to each one.

Particular nodes of the AMI application described in figure 1 have been added to our topology. Data collectors are deployed at neighborhood area network (NAN) while meter data management systems (MDMS) and distribution management system (DMS) were added at wide area network (WAN). To sum up the smallest branch of our topology counts a total of 14 nodes: 10 smart meters, 1 transformer, 1 substation, 1 MDMS and 1 DMS.

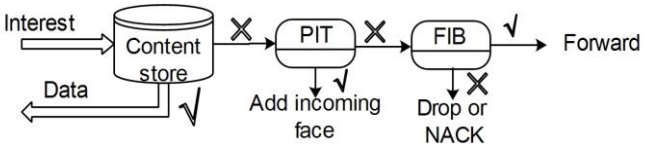


Fig. 2. Downstream interest processing in NDN

C. Simulation scenarios

Ndnsim used to simulate content-centric networking implements Named Data Networking [21] components under NS-3. It reproduces all its important building blocks: NDN core protocol, pending interest table (Pit), forwarding information base (Fib), Content Store (Cs) and applications. The following scenario exposes their interaction to manage communications. From a content-centric point of view, exchanging content occurs in two steps: a publisher producing content and a subscriber expressing interest in receiving it. Each network node has three data structures: a content store used for in-network caching, a pending interest table used to store received interests if the node can't satisfy them by a cache hit and a forwarding information base that stores possible forwarding paths to process incoming interests. When an interest is received, the Content Store (CS) is firstly checked. If the desired content is stored locally in the CS, then the node send it through the interest incoming face otherwise the Pending Interest Table is checked. If an entry with the same interest is found in the PIT, the interest incoming face is concatenated to existing faces and the interest itself is dropped. Interest existence in the PIT means that the node already received an interest in retrieving the same content and is waiting for a response. Once the desired content is received, it will be send to each face in the corresponding PIT entry. The final step is to check the Forwarding Information Base (FIB); it is carried out only if the interest does not exist in the PIT. Two scenarios are possible. Whether we find a corresponding entry for the interest which means that the node will forward it to the nodes found in the FIB and waits for them to send back the desired content. Or no entry is found in the FIB, this is the worst case since it means that the node has no idea how to find the corresponding content and will simply drop the interest (see figure 2). The upstream data processing is illustrated by figure 3. A node receiving a data packet first checks its content store to execute the caching policy. Of course whether to cache the content or not depends on the adopted caching placement strategy. For instance, if LCE (Leave Copy Everywhere) strategy is chosen, every node receiving a data packet appends it to its own CS. A replacement policy is also needed to define the way a CS is refreshed when the threshold size is reached. This policy describes which content to replace by the new incoming one; note that various replication and replacement strategies have been developed above in part III.

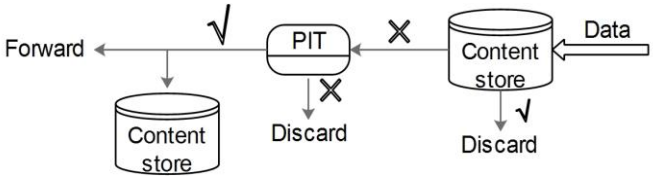


Fig. 3. Upstream Data processing in NDN

To assess in-network caching performances for AMI service, we deployed consumer and producer applications in the aforementioned topology nodes in order to simulate AMI data traffic described in Table 1. The first simulation scenario

aims at globally evaluating caching benefit for AMI. To do so, metrics were compared after enabling and disabling the caching functionality. Further simulation scenarios dig deeper into caching replacement policy and cache replication strategy evaluation. Simulations were run for 7200 seconds allowing to all data traffic to occur and in-network caching algorithms (placement and replacement) to take effect.

D. Simulation Results

1) In-network caching impact on global performances

The results of the first simulation show a reduced communication delay at each node of the topology and a reduced throughput (see fig 4). We remind that in this first simulation, the same AMI data traffic has been generated and two scenarios were compared; first with disabled in-network caching and second after enabling this feature. The delay represents the amount of time needed to satisfy an interest, it includes the queuing delay and the propagation delay. The reduction of communication delay is due to faster retrieval of data packets since closer nodes are satisfying data requests thanks to their content stores. The throughput is reduced significantly in comparison to that of the case without in-networking caching. Indeed less packets are travelling toward and backward source nodes since many interests are being served by content store hits.

2) Replacement strategy evaluation

After globally noticing in-network caching benefit while managing advanced metering communications, our goal is furthermore to investigate replacement strategy for AMI. Metrics evaluated at this level are:

- Hop count: the number of hops a data packet needs to travel in order to reach its requester.
- Cache hits: specifies the number of interests that were satisfied from the content store (CS).

Simulations were run using the same AMI topology and data traffic described in previous sections. Three replacement strategies are proposed by NDNSIM to insert new content packets once the content store size has been reached: FIFO (First In First Out), LRU (Least Recently Used) and LFU (Least Frequently Used). We vary cache size and notice that cache hits is proportional to cache size. It is a rational result since nodes with bigger cache size are able to store more data packets allowing, then, to satisfy more interests from CS. This fact is the reason behind the hop count decrease as the CS size increases. Indeed, as more interests are served from caches, shorter paths are taken by data packets to reach their requesters. Regarding the eviction strategies comparison, FIFO is having the lowest interest number satisfied from content stores (see fig 5). As for LRU and LFU, they are closely competing with a slight advantage for LFU strategy. On the other hand, the highest hop count reduction is observed with the LRU replacement strategy leaving FIFO and LFU behind. We remind that, if required, LRU removes content on recency of use basics; most recently used content is kept longer in CS. As for LFU, popularity of retrieval is what governs packet data eviction. We notice also, based on the curve allure, that the effectiveness of replacement strategies is more obvious when the content store (CS) size is beneath 50 packets. Thus, high content store size (beyond 50 packets) dissolves the replacement strategy impact on cache hits and hop count. Indeed, data packets won't get evicted from the CS unless storage threshold reached.

3) Placement strategy evaluation

The last stage of our simulation study aims at assessing various placement strategies. Placement strategy refers to a set of rules deciding where to cache content. In our context, four placement policies have been considered:

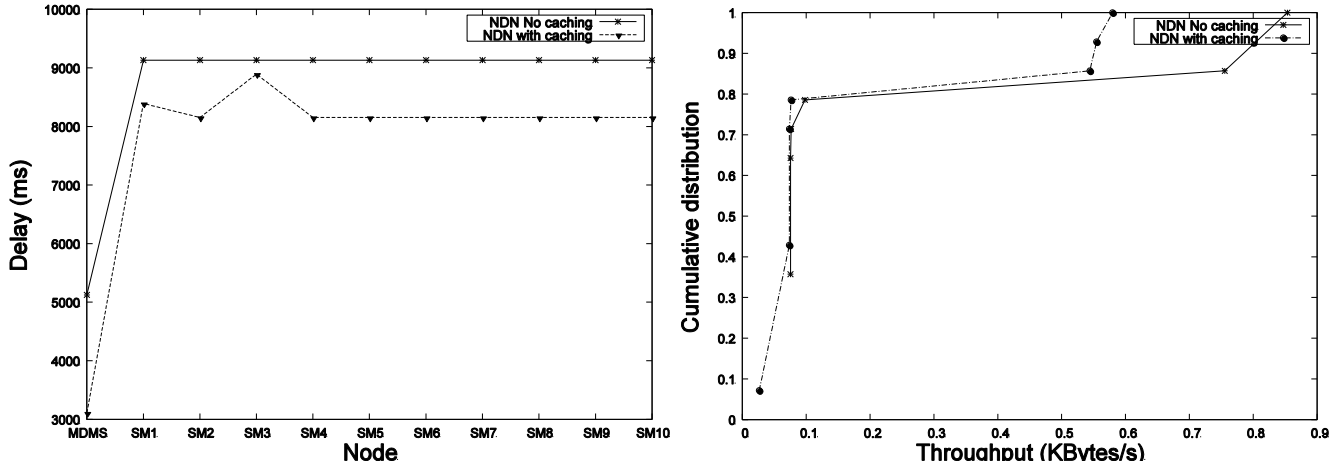


Fig. 4. Caching impact on delay and throughput

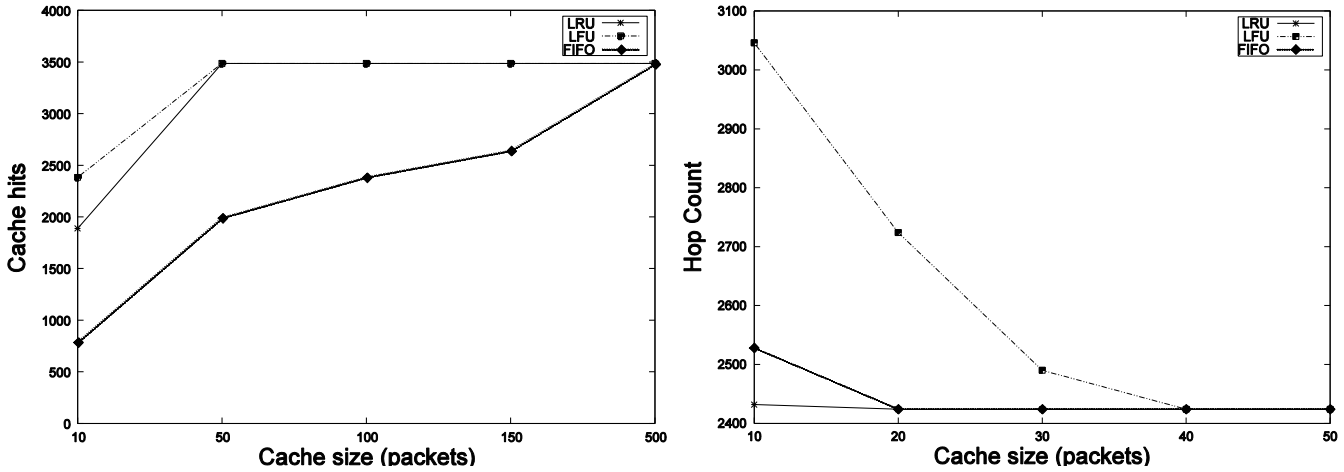


Fig. 5. Cache replacement policy evaluation

- LCE (Leave Copy Everywhere): this strategy allows the insertion of content in every node along its path toward its requester. Each content packet will be then stored in the content store of all nodes crossed while traveling back to the consumer.
- Probabilistic placement strategy: A normal distribution decides whether a content is stored in the content store or not.
- Probcache placement strategy [22]: Content is cached with a probability depending on two factors: remaining caching capacity on the downloading path and cache weight depending on content closeness to its requester. Probcache is considered as a cooperative caching mechanism since caching decision depends on factors provided by all routers along the path.
- Leave copy in none-constrained devices (LCNCD): Since we are assessing caching for a particular smart grid service, we take into consideration, in this placement strategy, the fact that smart meters are constrained devices. In this last placement strategy we only deploy content stores in none-constrained devices which exclude smart meters from in-network caching.

Fig.6 illustrates the results obtained after running the simulations described above with wider topology size. We varied branches number of the tree topology shown in fig.1 to reach a maximum of 5 data collectors and 50 smart meters. We observe a trivial impact of the cache placement strategy on the cache hits. The incidence on hop count is however more significant, especially with an important disadvantage of the probabilistic placement strategy that needs more hop number to satisfy interests. The most substantial result at this level is revealed by the superposition of LCE and LCNCD (Leave Copy in None-Constrained Devices) curves which imply that disabling the caching feature in the smart meters does not degrade the system performances. This result is due to the tree topology used in our simulations. Positioning smart meters at the edge of the network alters caching usefulness at their level. Probabilistic placement strategy is also less beneficial from delay point of view. Fig.7 shows, in fact, high packet transmission delays when normal probabilistic distribution governs cache placement. It strengthens, also, the fact that disabling cache in constrained field devices doesn't affect network performance as the transmission delay is not influenced with LCNCD placement policy.

VI. Conclusion

Advanced metering infrastructure in smart grids is responsible for sensing, measuring, collecting and sending consumption data and electrical parameters. This key role brings up AMI design weight on smart grid effectiveness. In the meanwhile, content-centric networking with its ubiquitous in-network caching feature is increasingly gaining attention as the future networking trend. It's premised that caching content on the delivery path can improve data delivery. Our goal was to investigate this assumption in handling AMI communication requirements. A simulation-based analysis reproducing AMI communication topology and data traffic showed reduced data delivery delay once in-network caching activated. It showed also inadequacy of FIFO replacement strategy for this context and better performances for content-popularity based strategies. Placement strategies analysis allowed us to show that disabling caching in smart meters does not affect the SGCS performance. It consolidates then our position on the adequacy of CCN to smart grids. This work concludes our series of articles investigating content-centric networking for smart grid communications. Our perspectives are the design of a content-centric smart grid communication infrastructure providing SGCS requirements such as latency, data transmission delay, quality of service and interoperability.

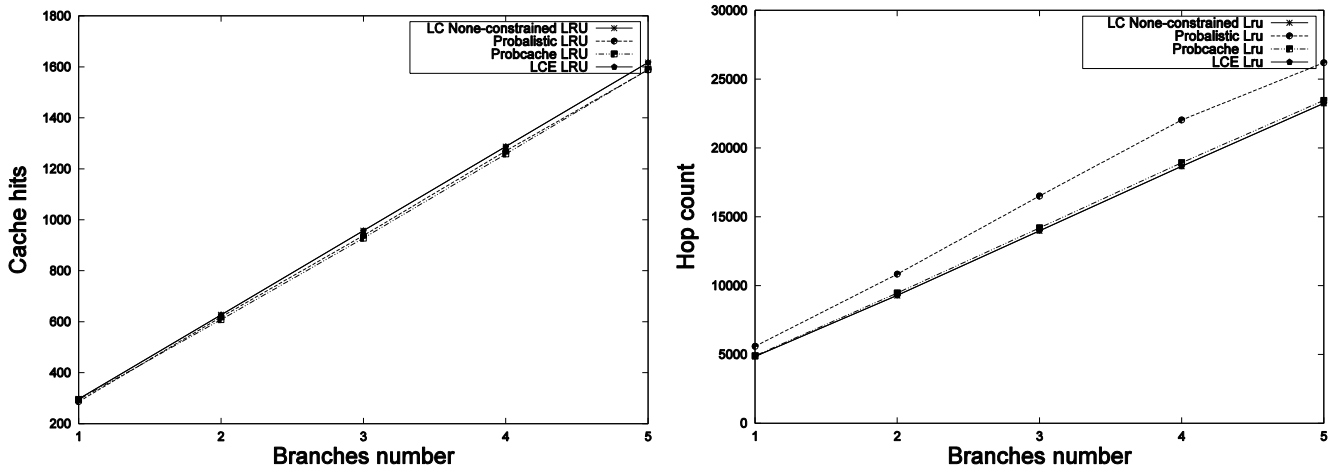


Fig. 6. Cache placement policy evaluation

REFERENCES

[1] N. E. H. BenYoussef, Y. Barouni, S. Khalfallah, J. B. H. Slama, and K. B. Driss, "Evaluating content-centric communication over power line communication infrastructure for smart grids," *Procedia Computer Science*, vol. 73, pp. 217 – 225, 2015. [Online]. Available: http://-www.sciencedirect.com/science/article/pii/S1877050915034766

[2] "Supporting renewable energy resources integration using content-centric networking," in *Networks, Computers and Communications (ISNCC), 2016 IEEE 3rd International Symposium on*, May 2016.

[3] K. Yu, L. Zhu, Z. Wen, A. Mohammad, Z. Zhou, and T. Sato, "Ccn-ami: Performance evaluation of content-centric networking approach for advanced metering infrastructure in smart grid," in *Applied Measurements for Power Systems Proceedings (AMPS), 2014 IEEE International Workshop on*, Sept 2014, pp. 1–6.

[4] K. Young-Jin, L. Jaehwan, G. Atkinson, K. Hongseok, and M. Thottan, "Sedax: A scalable, resilient, and secure platform for smart grid communications," *IEEE Journal on Selected Areas in Communications*, vol. 30, no. 6, pp. 1119–1136, 2012.

[5] K. Katsaros, W. C., N. W., G. Pavlou, H. Bontius, and M. Paolone, "Information-centric networking for machine-to-machine data delivery: a case study in smart grid applications," *IEEE Network*, vol. 28, no. 3, pp. 58–64, May 2014.

[6] W. K. Chai, N. Wang, K. V. Katsaros, G. Kamel, G. Pavlou, S. Melis, M. Hoefling, B. Vieira, P. Romano, S. Sarri, T. T. Tesfay, B. Yang, F. Heimgaertner, M. Pignati, M. Paolone, M. Menth, E. Poll, M. Mampaey, H. H. I. Bontius, and C. Develder, "An information-centric communication infrastructure for real-time state estimation of active distribution networks," *IEEE Transactions on Smart Grid*, vol. 6, no. 4, pp. 2134–2146, July 2015.

[7] J. Zhang, Q. Li, and E. Schooler, "ihems: An information-centric approach to secure home energy management," in *Smart Grid Communications (SmartGridComm), 2012 IEEE Third International Conference on*, Nov 2012, pp. 217–222.

[8] G. Piro, I. Cianci, L. Grieco, G. Boggia, and P. Camarda, "Information centric services in smart cities," *Journal of Systems and Software*, vol. 88, pp. 169 – 188, 2014. [Online]. Available: http://-www.sciencedirect.com/science/article/pii/S0164121213002586

[9] R. R. Mohassel, A. Fung, F. Mohammadi, and K. Raahemifar, "A survey on advanced metering infrastructure," *International Journal of Electrical Power & Energy Systems*, vol. 63, pp. 473 – 484, 2014. [Online]. Available: http://www.sciencedirect.com/science/article/pii/-S0142061514003743

[10] D. F. Ramírez, S. Céspedes, C. Becerra, and C. Lazo, "Performance evaluation of future ami applications in smart grid neighborhood area networks," in *Communications and Computing (COLCOM), 2015 IEEE Colombian Conference on*, May 2015, pp. 1–6.

[11] F. Aalamifar, "Viability of powerline communication for smart grid realization," Master's thesis, Queen's University, 2012.

[12] T. Koponen, M. Chawla, B. Chun, A. Ermolinskiy, K. H. Kim, S. Shenker, and I. Stoica, "A data-oriented (and beyond) network architecture," *SIGCOMM Computer Communication Review*, vol. 37, no. 4, pp. 181–192, Aug. 2007.

[13] M. Gritter and D. R. Cheriton, "An architecture for content routing support in the internet," *Proceedings of the 3rd Conference on USENIX Symposium on Internet Technologies and Systems - Volume 3*, pp. 4–4, 2001.

[14] H. Balakrishnan, K. Lakshminarayanan, S. Ratnasamy, S. Shenker, I. Stoica, and M. Walfish, "A layered naming architecture for the internet," *SIGCOMM Comput. Commun. Rev.*, vol. 34, no. 4, pp. 343–352, Aug. 2004.

[15] V. Jacobson, D. K. Smetters, J. D. Thornton, M. F. Plass, N. H. Briggs, and R. L. Braynard, "Networking named content," *Proceedings of the 5th International Conference on Emerging Networking Experiments and Technologies*, pp. 1–12, 2009.

[16] M. Zhang, H. Luo, and H. Zhang, "A survey of caching mechanisms in information-centric networking," *IEEE Communications Surveys Tutorials*, vol. 17, no. 3, pp. 1473–1499, thirdquarter 2015.

[17] G. Xylomenos, C. Ververidis, V. Siris, N. Fotiou, C. Tsilopoulos, X. Vasilakos, K. Katsaros, and G. Polyzos, "A survey of information-centric networking research," *Communications Surveys Tutorials, IEEE*, vol. PP, no. 99, pp. 1–26, 2013.

[18] I. Abdullahi, S. Arif, and S. Hassan, "Survey on caching approaches in information centric networking," *J. Netw. Comput. Appl.*, vol. 56, no. C, pp. 48–59, Oct. 2015. [Online]. Available: http://dx.doi.org/10.1016/-j.jnca.2015.06.011

[19] A. Ioannou and S. Weber, "A survey of caching policies and forwarding mechanisms in information-centric networking," *IEEE Communications Surveys Tutorials*, vol. PP, no. 99, pp. 1–1, 2016.

[20] Y. Ye, Q. Yi, H. Sharif, and D. Tipper, "A survey on smart grid communication infrastructures: Motivations, requirements and challenges," *Communications Surveys Tutorials, IEEE*, vol. 15, no. 1, pp. 5–20, 2013.

[21] L. Zhang, A. Afanasyev, J. Burke, V. Jacobson, k. claffy, P. Crowley, C. Papadopoulos, L. Wang, and B. Zhang, "Named data networking," *SIGCOMM Comput. Commun. Rev.*, vol. 44, no. 3, pp. 66–73, Jul. 2014. [Online]. Available: http://doi.acm.org/10.1145/2656877.2656887

[22] I. Psaras, W. K. Chai, and G. Pavlou, "Probabilistic in-network caching for information-centric networks," in *Proceedings of the Second Edition of the ICN Workshop on Information-centric Networking*, ser. ICN '12. New York, NY, USA: ACM, 2012, pp. 55–60. [Online]. Available: http://doi.acm.org/10.1145/2342488.2342501

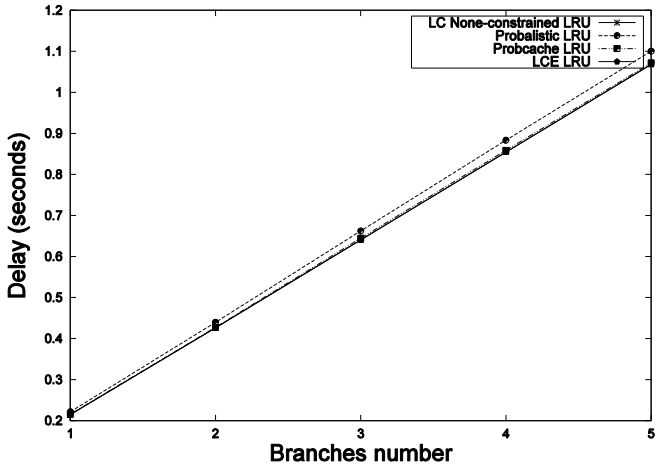


Fig. 7. Impact of placement strategy on transmission delay

Development of Dynamic Real-Time Navigation System

Shun FUJITA
Graduate School Student,
Graduate School of Information Systems,
University of Electro-Communications
Tokyo, Japan

Kayoko YAMAMOTO
Associate Professor,
Graduate School of Informatics and Engineering
University of Electro-Communications
Tokyo, Japan

***Abstract*—This study aimed to develop a system that considers dynamic real-time situations to provide effective support for tourist activities. The conclusions of this study are summarized in the following three points: (1) The system was developed by integrating Web-GIS, social media, recommendation systems and AR terminals (smart glasses) into a single system, and operated in the center part of Yokohama City in Kanagawa Prefecture, Japan. It enabled the accumulation, sharing and recommendation of information and navigation to guide users to their goals both in normal conditions and in the event of disasters. (2) The web-based system was aimed at members of the general public over 18 years old and operated for seven weeks. The total number of users was 86, and 170 items of information were contributed. A system using smart glasses operated for two days, and the total number of users was 34. (3) Evaluation results clarified that it was possible to support user behavior both in normal conditions and in the event of disasters, and to efficiently and safety conduct navigation using smart glasses. Operation premised on disaster conditions showed that users who accessed the system via mobile information terminals increased, and actively used functions requiring location information.**

***Keywords*—Navigation System; Dynamic Real-Time; Web-Based Geographical Information Systems (GIS); Social Media; Recommendation System; Augmented Reality (AR); Smart Glasses**

I. INTRODUCTION

As many who visit sightseeing spots do not have a good sense of locality, they get to know the route to their destination by means of guidebooks using paper maps. However, carrying around maps is inconvenient, and because users must look up the route to their destination by constantly checking their current location on the map, this time consuming process may reduce their desire to sightsee. In addition, because many tourists do not have any knowledge about local disaster countermeasures including evacuation sites and support facility locations in the event of disasters, it will be extremely difficult for them to take necessary actions to evacuate. Also, in the event of disasters, although there are systems for supporting the evacuation of residents in the affected area, as these disaster countermeasure systems are not used in normal conditions, it will be difficult to suddenly use this system when disasters actually occur. Therefore, a system that is used in normal conditions in addition to one that supports the evacuation in the event of disasters by means of the same method used for normal conditions is necessary. From what is mentioned above, by means of the information system using the information of the situation around the user's location, in addition to appropriate support of sightseeing and evacuation, users can sightsee more efficiently and safely than before.

On the other hand, although navigation systems using mobile information terminals including smartphones are often used in recent years, using a smartphone while walking is called "wexting", and can be dangerous as it makes it hard for users to grasp their own surroundings. In contrast, with navigation using AR terminals (smart glasses) which is a type of wearable terminal, as information is displayed in front of the user's eyes without any special process, it is easy to grasp one's surroundings and can help users navigate safely. In addition, with the spread of social media in recent years, information related to sightseeing and disasters are being submitted and updated on social media in real time. Therefore, gathering real time information through social media, and reflecting that information both in sightseeing support in normal conditions and evacuation support in the event of disasters is necessary to realize a more efficient and safe sightseeing environment than before. Based on the circumstances mentioned above, the purpose of this study is to develop a navigation system that can actively alter routes, in order to gather high real time information concerning urban tourist spots as well as support sightseeing in normal conditions as well as evacuation in the event of disasters.

II. RELATED WORK

This study is related to (1) the study concerning the sightseeing support system, (2) the study concerning the Point of Interest (POI) recommendation system, and (3) the study concerning social media GIS. To list the representative preceding studies in recent times of the three groups mentioned above, regarding (1), Kurata (2012) [1] developed an interactive trip planning support system using genetic algorithm which can be used on the web. Sasaki et al. (2013) [2] developed a system that collects information regarding regional resources, and supports the tours of each user. In addition, Fujitsuka et al. (2014) [3] developed an outing plan recommendation system using the pattern mining method that lists and extracts the time series activities of users visiting sightseeing spots. Ueda et al. (2015) [4] generated post-activity information from the user's activities while sightseeing, and developed a sightseeing support system that shares the information as pre-activity information for other users.

Regarding (2), Noguera et al. (2012) [5] developed a POI recommendation system by means of the methods of both collaborative recommendation and knowledge-based recommendation on 3D maps, using mobile terminals with the location information as the basis. In addition, studies related to POI recommendations concerning LBSN (Location-Based Social Networks) are also included in this field. Meo et al. (2011) [6] proposed and evaluated the preference and social and geographical effect of users concerning LBSN, Yuan et al. (2013) [7] proposed and evaluated the time and space information using the check in data concerning LBSN, and Chen et al. (2016) [8] proposed and evaluated the POI recommendation method which considers the interrelationship between users concerning LBSN.

Regarding (3), Yanagisawa et al. (2011) [9] in addition to Nakahara et al. (2012) [10] developed an information sharing GIS with the purpose of accumulating and sharing information regarding the local community using Web-GIS, SNS and Wiki. Yamada et al. (2013) [11] and Okuma et al. (2013) [12] developed a social media GIS which reinforced the functions of social media included in the information sharing GIS as mentioned above. By using the systems of these preceding studies as a base, Murakoshi et al. (2014) [13] in addition to Yamamoto et al. (2015) [14] developed social media GIS for the utilization support of disaster information assuming it will be used continuously from normal conditions to when disasters occur. Additionally, with the social media GIS as a base, Ikeda et al. (2014) [15] developed social recommendation GIS which recommends sightseeing spots according to the preferences of each user by integrating the recommendation system with the social media GIS mentioned above.

Among the preceding studies in related fields as listed above, (1)(2)(3) support the tour planning and accumulating, sharing and recommending of sightseeing spot information for sightseeing activity support in normal conditions, and (3) performs the accumulating and sharing of disaster information for evacuation support in the event of disasters. However, these preceding studies do not go further than offering information to users by means of accumulating, sharing and recommending information, which is realistically not enough to support the activities of users. Additionally, they do not support the users' activities for both sightseeing in normal conditions and evacuation in the event of disasters. In contrast to these preceding studies in related fields, this study shows individuality in developing a system, which supports sightseeing in normal conditions and evacuation in the event of disasters by means of navigation, by integrating SNS, Twitter, Web-GIS, recommendation systems and smart glasses.

III. Outline and Method of this Study

This study will follow the outline and methods as shown below. First, the navigation system which specializes in the purpose of this study will be designed (Section III) and developed (Section IV) originally. Next, assuming users are over the age of 18, operation tests and operation (Section V) of the navigation system in addition to the evaluation and extraction of solutions (Section VI) will be conducted. Also, assuming each user will be using this system for approximately 1 month, operation will start after conducting operation test and evaluations. In addition, Web questionnaire surveys to users and access analysis using the log data during the operation will be conducted, and with the obtained results, the solution extraction for this system will be conducted by evaluating the system.

The central part of Yokohama city in Kanagawa Prefecture, Japan has been selected as the region of operation. The reason for this is that (1) there is a variety of sightseeing spots, as it is an urban tourist destination, which enable recommendations of sightseeing spots to be made according to each user's preferences, and (2) because tourists and sightseeing spots are concentrated in a small area, a lot of information concerning sightseeing spots are transmitted and the obtainment of real time information is made possible.

IV. System Design

A. System configuration

The system of this study is developed by means of SNS, Twitter, Web-GIS, recommendation system and smart glasses, as shown in Fig. 1. With the support of both sightseeing and evacuation as the purpose of this study, the activities of users are supported by means of the navigation that can actively change routes. Concerning navigation, in the case of sightseeing, information of the route to a single sightseeing spot from the current location of the user as well as routes for touring a group of sightseeing spots are provided. Additionally, in the case of evacuation, information including the closest evacuation sight from the current location of the user and routes to the support facility in the event of disasters are provided. Also, on Twitter and SNS, which is originally developed, information concerning sightseeing spots including that of various events as well as disaster information of nearby areas will be obtained in real time, and the navigation route will change according to the information provided if necessary. By means of such a navigation system that focuses on obtaining real time information, the efficient support of both sightseeing and evacuation is realized.

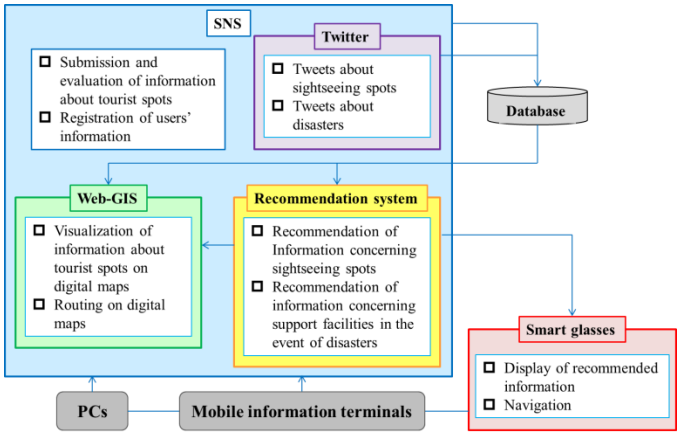


Fig. 1. System design of dynamic real-time navigation system

B. System usability

The usability of this system can be described in detail as shown below by means of the design mentioned in the previous section.

(1) Relaxation of time restrictions

Concerning PCs and mobile information terminals, the system gets a time restriction in providing information to users, as it provides information after receiving requests from users,. By using smart glasses, the information provided by the system will constantly be displayed in front of the users. Therefore, because information is provided regardless of whether or not users make requests, the system will not have to rely on time restrictions. Consequently, it will be possible for information to be provided to users without losing its real time quality.

(2) Active and real time information

Without being limited to information accumulated in the past, by immediately reflecting real time information, which is obtained through originally developed SNS and Twitter, in recommendations, information can be provided to users efficiently. Based on such real time information, active information appropriate to each user's location can be provided.

(3) Load reduction of information obtainment

In the case of information overload along with the system's long-term operation, this system can provide information appropriate for each user by means of the recommendation system. In addition, as information is provided whenever necessary while using the smart glasses, opportunities for users to use information terminals in order to request information will be reduced. Therefore, the load of users when obtaining information can be reduced, and sightseeing in normal conditions and evacuation in the event of disasters can be focused on.

C. Target terminals

This system is set with the assumption that it will be used from PCs, mobile information terminals and smart glasses. Because PCs are assumed to be used indoors, the submitting, viewing and recommending function of sightseeing spot information, registration function of activity history, support function for planning sightseeing trips and navigation functions, which are described in detail in Section IV, will all be available. Assuming mobile information terminals will be used indoors and outdoors, the submitting and viewing function of sightseeing spot information as well as the navigation function will be the main functions. However, because using mobile devices while walking can be dangerous, assuming the use of the smart glasses which displays information in front of the user as well, safe navigation will also be realized.

D. System operating environment

This system uses the web server, database server and the GIS server for operation. The web server and database server were prepared using the Heroku. Heroku is the PaaS supplied by the Salesforce company, and it provides a platform which operates web applications. In addition, GIS servers use the ArcGIS Online of the ESRI. As the main language, the web applications developed by this system are implemented by Python and JavaScript, while the smart glasses' application is implemented by Java.

E. System structure

1) SNS

In this study, an original SNS, which can be integrated with Twitter, Web-GIS, the recommendation system and smart glasses, will be designed. As the purpose of the designed SNS will not only be communication between users but also the gathering of sightseeing spot information, which will serve as the base information for preference information of users as well as recommendations, functions that promote friend registration and community communication will not be implemented. The main functions of the designed SNS will be the registration of user information and the submitting, viewing and recommending of information. The information made public through profiles will use nicknames instead of real names, and information that may identify an individual, such as gender or age, will not be made public. The comment function and tag function were designed as a method of communication. Comments that are submitted by means of the comment function will also be used as real time information. In addition, tags that have been added to sightseeing spots by means of the tag function will be treated as features of the sightseeing spots, and will be used by the recommendation function of sightseeing spots with the preference information of users as the base.

2) Web-GIS

As this study assumed that the system users will be an unspecified large number from both inside and outside the region of operation, it is better if the system can be used by means of a web browser instead of having users install a special software. Additionally, it is necessary to conduct route searches as well as information visualization on digital maps. Therefore, the Web-GIS, which was developed using the ArcGIS API for JavaScript of the ESRI, will be used. Also, for route searches, the ArcGIS Online Directions and Routing Services will be used. Although Google Maps was the most used in preceding studies of related fields listed in Section II, as searches of routes that do not go through a specific location, which is necessary for evacuation support in the event of disasters, are impossible concerning this system, ArcGIS API for JavaScript is used in this study.

The process up to the display of routes regarding Web-GIS, which used the ArcGIS API for JavaScript and ArcGIS Online Directions and Routing Services as mentioned above, is as shown in Fig. 2. First, the request for a route search as well as search criteria will be sent from ArcGIS for JavaScript API to the ArcGIS Online Directions and Routing Services. Next, after receiving route search results from the ArcGIS Online Directions and Routing Services, the route will be displayed on the Web-GIS developed by means of ArcGIS for JavaScript API.

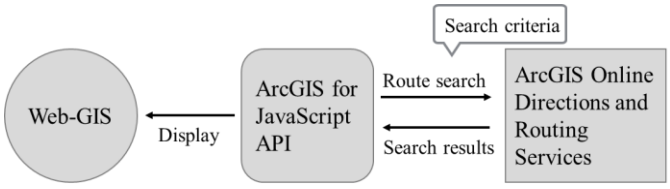


Fig. 2. Process up to the route display

3) Recommendation system

The recommendation system has 3 methods including the collaborative recommendation, the content-based recommendation and the knowledge-based recommendation (Jannach et al., 2012) [16]. The recommendations of this study, which is based on the preference information of users concerning the sightseeing support in normal conditions, are conducted using 2 methods. The first is the knowledge-based recommendation, which recommends sightseeing spots that have features which match the tags registered as preference information by users. The reason for using the knowledge-based recommendation is to solve the cold-start problem, which is when appropriate information to recommend new users to this system is lacking. Therefore, as it was decided that knowledge-based recommendation which explicitly ask users for preference information and create profiles are necessary, the system was originally designed according to the purpose of this study.

The second is the collaborative recommendation, which supports the tour planning based on the activity history registered by users. Activity history lists sightseeing spots, which were visited in one day, in chronological order, and shows whether users were satisfied or dissatisfied with each sightseeing spot. Based on this information, a user model is created using the Support Vector Machine (SVM), the model is matched with the activity history of other users, the degree of similarity is determined, and users with similar preferences are detected. Afterwards, regarding the activity history of users with similar preferences, a pattern mining, which lists and extracts time series activities of sightseeing a group of locations, will be conducted. The pattern mining in this study will use categories. The 7 categories, which divide sightseeing spots according to each feature, used in this study include food/drinks, shops, entertainment, events, scenery, art and recreation. All sightseeing spot information belongs to one of these categories. Therefore, each sightseeing spot information in the activity history will be put in the category it belongs to, and the categories corresponding to the list of sightseeing spots visited in chronological order will be used as the new activity history. Then, the system will extract patterns of categories from the activity history of users with similar preferences, and recommend patterns with the highest relativity score as described in Section V. Regarding each category within the recommended patterns, by selecting and matching sightseeing spot information belonging to each category, a tour plan is made. In addition, concerning evacuation support, by calculating the distance based on longitude and latitude, the closest evacuation site as well as disaster support facilities within a certain distance can be recommended based on the user's current location.

4) Smart glasses

Smart glasses are a wearable terminal in the form of glasses, and have been developed by various companies in recent years. In this study, the system will be designed with the assumption that the smart glasses made by SONY will be used. The reasons why the smart glasses was selected are because the binocular lenses provide high visibility in various environments, and by using the GPS of cooperating mobile information terminals, information according to the current location of users can be provided. In addition, concerning the Osaka marathon held in Osaka City on October 2014, from the demonstration experiment where the runners ran with the smart glasses on, it is anticipated that the products have been developed with the concept that it will be used while users are moving, which is compatible with the purpose of this study where it is assumed that the system will be used while users are walking.

With the smart glasses, the navigation is conducted by providing the direction and distance from the current location to the destination by means of the location information. In addition, while the navigation system is working, real time information concerning the destination will be provided whenever necessary. The information provided is information regarding sightseeing spots obtained by SNS comments and Twitter.

F. Management of submitted information

As the standard of validity of submitted information concerning sightseeing spots are vague, it is difficult to determine whether or not the content is appropriate. However, if information submitted by users are not managed, when a user with ill intent appears, the reputation of a specific sightseeing spot may be damaged arbitrarily. If users have knowledge concerning the region of operation, the legitimacy of submitted information can be determined based on experience. However, users who do not have any knowledge will not be able to determine the legitimacy and may make incorrect evaluation concerning sightseeing spots. In this case, the information of the sightseeing spot is inappropriate, and it will affect information recommendations as recommendations suitable for users will not be possible, which will ultimately damage the value of the system. Therefore, in this study, if the manager detects any submissions with ill intent, regarding the account that created the submission with ill intent, the system is designed so that authority to delete the submission and account can be exercised. By means of this, this study aims to design a system that can operate on a long-term basis.

V. System Development

A. System frontend

1) Functions for sightseeing support in normal conditions

a) Submitting function of sightseeing spot information

By clicking on the “spot submissions” in the menu bar, users will be moved to the submitting page of sightseeing spot information. On the submitting page of sightseeing information, users can submit sightseeing spot information by entering the name, description, images and location information of the sightseeing spot. The location information of the sightseeing spot can be entered by clicking the target location on the Web-GIS. In addition, by clicking the “display past submitted locations”, users can confirm whether the same sightseeing spot information has been submitted in the past.

b) Viewing function of sightseeing spot information

Users can return to the homepage by clicking “home” in the menu bar, and view sightseeing spot information submitted by users in the past on the Web-GIS. Each sightseeing spot information is displayed with different color markers according

to each category, and the category of each marker is explained in the image below the Web-GIS. When clicking the marker, a bubble containing the name and image of the sightseeing spot will be displayed. By clicking the bubble, users will be moved to the details page of the selected spot which will enable them to check the detailed information.

In the details page of sightseeing spots, the comment and tag functions can be used. The comment function will enable communication between users as well as supplementary information to be added to sightseeing spot information. In addition, Tweets related to sightseeing spots obtained through Twitter is also displayed in the comment section. These comments and Tweets are considered real time information of sightseeing spots. Regarding the tag function, the features of sightseeing spots can be freely added as tags by users. Users can use tags that have been registered in the past as well as tags that have been newly registered. Additionally, by clicking tags that have already been added, additional importance can be placed on the tag. The most commonly used tag will be used as the feature of the sightseeing spot when recommended. All tags belong to a category, and the category that the most-used tag of a sightseeing spot belongs to will also be the category in which the sightseeing spot belongs to. In addition, by clicking on "start navigation from your current location", users can receive navigation on the Web-GIS to any sightseeing spot from their current location.

c) Registering function of activity history

By clicking on "history" in the menu bar, users will be moved to the activity history registration page. Activity history is made up of users' evaluation of previously visited sightseeing spots, within the region of operation, in addition to the budget and group when visiting the sightseeing spots. The number of sightseeing spots that can be registered are 2 to 5. In addition, if the activity history is already registered, the confirmation screen of activity history will appear. On the confirmation screen of activity history, by clicking the button with each activity history name, users can confirm the activity history registered in the past. The contents of activity history that will be displayed include the name, image and category of each sightseeing spot.

d) Support function of tour planning

By clicking "plans" in the menu bar, users will be moved to the tour planning page. Concerning the tour planning page, users can receive tour planning support from the system based on the registered activity history. First, concerning the tour planning, the budget and group and the number of sightseeing spots that the users would like to visit must be entered and sent as conditions. Based on the conditions and activity history of each user, the system will recommend patterns made up by categories. Regarding each category in the pattern recommended, users can select and match sightseeing spots belonging to each category and create a tour plan. In addition, in order to efficiently create a tour plan, the location information of sightseeing spots applied to each category will be actively displayed on the Web-GIS. If the tour plan is already made, users will be moved to the tour plan confirmation screen, and the name, image and category of sightseeing spots will be displayed according to the order on the sightseeing schedule. To send the displayed tour plan to the smart glasses, users must click the "setup the plan in smart glasses". By clicking the "start navigation", the navigation of the tour plan will start on the Web-GIS.

e) Navigation function

By clicking the "start navigation from current location" on the details page of sightseeing information or the "start navigation" on the confirmation screen of the tour plan, the first option will take the users to the navigation screen for single sightseeing spots, and the second will take users to the navigation screen for several sightseeing spots. Regarding navigation for single sightseeing spots, navigation will be conducted by displaying the current location and the route from the current location to the user's destination on the Web-GIS. In addition, for navigation of several sightseeing spots, the navigation will be conducted by simultaneously displaying the current location and the route for several sightseeing spots.

f) Recommendation function of sightseeing spot information

Users will be moved to the recommendations page by clicking the "recommended" in the menu bar, and sightseeing spots with the most-used tags that are also registered as the users' preference information will be recommended. The information of the recommended sightseeing spots will be listed in tile form, and the content will include the name, description, image and category of each sightseeing spot.

2) Functions for evacuation support in emergency situations in the event of disasters

a) Viewing function of support facilities in the event of disasters

Users can go to the homepage by clicking "home" in the menu bar, and view information of disaster support facilities (evacuation locations, evacuation sites, temporary accommodation, water supply points and medical institutions), published by the disaster prevention map of Yokohama city which is within the region of operation, on the Web-GIS. The information of these disaster support facilities are marked differently according to the facility on the digital map of the Web-GIS, and the type of marked facilities are explained below the image of the Web-GIS. When clicking the marker, a bubble with the name of the disaster support facility will be displayed, and for more detailed information, users can click the bubble which will take them to the details page of the selected disaster support facility. The comment function can be used on the details page of disaster support facilities. The comment function enables communication between users and the supplementation of information concerning the disaster support facilities. In addition, by clicking the "start navigation from current location", users can receive navigation from their current location to the selected disaster support facility on the Web-GIS.

b) Search function of support facilities near users in the event of disasters

Users can go to the nearby disaster support facility page by clicking the "nearby disaster support facilities" in the menu bar, and disaster support facilities that are within 1km from the user's current location will be recommended. The

recommended disaster support facilities will be listed in tile form, and the name, facility type and distance from the user's current location to the facility of each disaster support facility will be displayed.

c) Navigation function to the closest evacuation sites

By clicking the "evacuation navigation" in the menu bar, users can move to the navigation page that shows the route to the closest evacuation location from the user's current location.

B. System backend

1) Information obtained through Twitter

In this study, real time information is gathered through Twitter in addition to the originally designed SNS. Twitter API 1.1 is used to obtain Tweets from Twitter. Tweets that are obtained in normal conditions are those submitted within 24 hours that include the words "Minatomirai" or "#Minatomirai", and have been retweeted or added to favorites 5 times each. Taking the text out of each obtained Tweet, a search of letter strings by means of regular expressions will be conducted. The searched letter string is to be the names of all sightseeing spots submitted on SNS, and if a matching letter string is found, that Tweet will be registered in the database as comment relating to the sightseeing spot matched by the letter string. In the event of disasters, Tweets including "#DynamicNavigation" in addition to having location information will be obtained. This system obtains Tweets every minute because the Twitter API has restricted the number of Tweet obtainment to 15 times in 15 minutes.

2) Calculation of the similarity ratio of preferences between users

With the method of Fujitsuka et al. (2014) [2] as a reference, SVM will be used for the similarity rate calculations of preferences between users in this study. SVM is one of the methods of machine learning, and it can also make models that discern different patterns in order to divide data into several classes. In this study, users' activity history will be used as learning data, and budget/group and tags will be treated as features, while class will be divided into satisfactory class and dissatisfactory class. First, a user model based on the activity history of users will be made, the activity history of a different user will be applied to it, and the data will be divided into the satisfactory class or the dissatisfactory class. By comparing the aforementioned user model and the separation results, the ratio of matched satisfaction and dissatisfaction among the activity history will be set as the preference similarity rate between users. Using the study results of Fujitsuka et al. (2014)[2] as a reference, and basing the level of similarity rates discerned as being able to appropriately determine similar users by the operation tests (Section VI) mentioned later, this study will set those with a similarity rate of over 60.0% as users with similar preferences.

3) Creating recommendations

In normal conditions, in order to support the sightseeing spot recommendation based on users' preference information as well as the tour planning based on users' activity history, this system has made recommendations regarding users. Concerning the former, tags registered as user preference information and the most-used tags among the tags included in sightseeing spot information will be put together, and the matching sightseeing spot information will be recommended. Concerning the latter, calculations of the similarity rate between users will be made, and based on the activity history of users with similar preferences as well as the conditions presented from users, recommendation will be made using the pattern mining method. In particular, sightseeing spot information included in the activity history will first be converted to the category it belongs to, and the category group will be made. Then, all patterns from the category groups will be extracted. Next, as shown in Fig. 3, the closeness (degree of distance) will be solved with the chronological order of the categories in mind in contrast with the extracted patterns, the closeness score of each pattern will be calculated by multiplying the support rate of patterns (appearance ratio), and this will enable the recommendations with the highest pattern. Also, in the event of disasters, based on the current location information of users, the closest evacuation location and disaster support facilities within a certain distance from the users will be recommended.

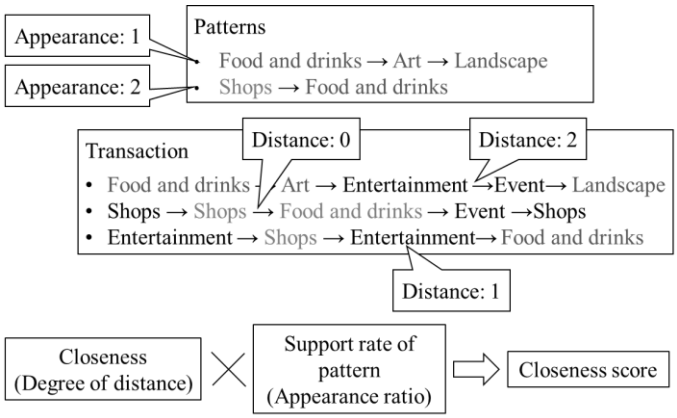


Fig. 3. Calculation method of the adjacency ratio score

4) Change of route

During navigation, by means of comments submitted on SNS and information obtainment through Twitter, routes will be changed when the system perceives higher real time information. Concerning normal conditions, when comments and Tweets concerning sightseeing spots are registered to the system, users receiving navigation will be notified about this and whether the user would like to visit the sightseeing spots. If the user would like to visit the sightseeing spot, the route will change by adding the sightseeing spot. Concerning disasters, if Tweets concerning the disaster which also include the word "DynamicNavigation" and the location information are obtained, the location information will be extracted from the Tweet, and it will be dealt as barrier information. Afterwards, by searching for new routes that arrive at the destination without going through the location shown in the extracted barrier information, the route will be changed.

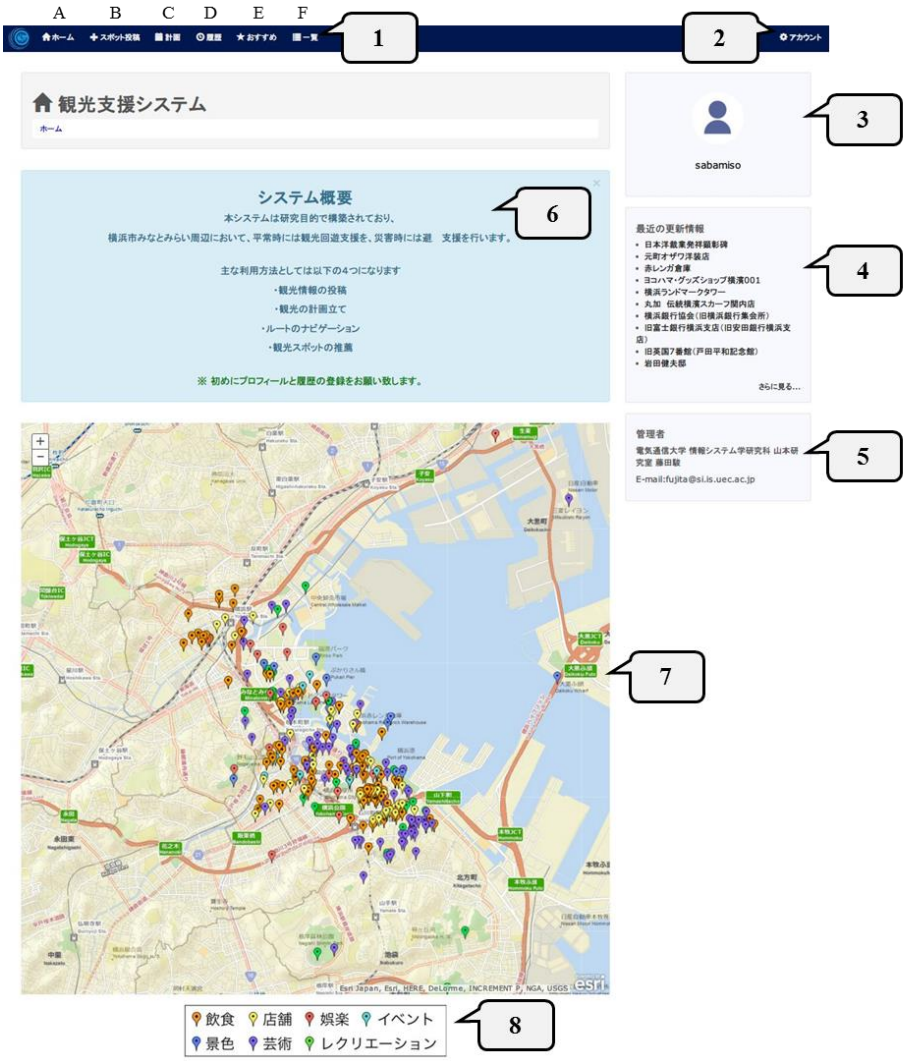
5) Switching to the emergency mode

This system can be switched to emergency mode only by the manager in the event of disasters. Concerning the switch to emergency mode, the manager can either rewrite the text file within the system, or click the "switch to emergency mode" button which only appears on the home screen of the manager

after logging in. In addition, the switch to normal mode from emergency mode can be done in the same way only by the manager.

C. System interface

The interface is optimized according to the user's PC screens (Fig. 4), mobile information terminal screens (Fig. 5), smart glasses screens (Fig. 6) and manager screens. The PC screen interface has the layout with a menu bar allowing easy access to each function. In addition, because it is designed to use 1 function on 1 page, users who are new to the system can easily use it. The interface for mobile information terminals is basically the same as PCs, but by changing the layout and size of items according to the size of the screen, the operability of the system is made easy. Regarding the interface for smart glasses screens, the distance and direction of the destination from the user's current location will be displayed and comments and Tweets concerning the destination will also be provided when necessary. Also, in order to maintain safety while users are walking, information will be displayed only on the bottom half of the screen. Managers can manage information saved in the database, which include personal information of users' and submitted information, on the manager screen. Information is displayed in a list form on the manager's screen, and as information is deleted using the Graphical User Interface (GUI) operation, the system is designed so that management can be possible regardless of the manager's IT literacy.



No.	Description
1	A: Go to initial page B: Submitting function of sightseeing spot information C: Support function of tour planning D: Registering function of activity history E: Recommendation function of sightseeing spot information F: Viewing function of sightseeing spot information
2	Go to user information change and registration page
3	Display of user information
4	Display of ten latest pieces of submitted tourist spot information
5	Display of manager information
6	Explanation about this system
7	Display of tourist spot information on digital map
8	Marker legend

Fig. 4. Interface for PCs (Normal mode assuming normal conditions)

No.	Description
1	Go to initial page
2	Display of tourist spot information on digital map
3	Marker legend

Fig. 5. Interface for mobile information terminals (Normal mode assuming normal conditions)

Fig. 6. Image of the interface for the smart glasses

VI. Operation Tests and Operation

Following the operation process in TABLE I, the operation was conducted after the operation test and operation test evaluation of the navigation system that was designed and developed in this study.

A. User assumption

Before the operation, 6 students in their 20s were selected and a two-week operation test was conducted. From the hearing survey results of the test subjects after the operation test, three points of improvement were found in areas including the location information of tourist spots on the detailed information viewing page, the information display method on an AR Smart Eyeglass, and the updating of tourist spot information displayed on Android apps, and the system was restructured regarding these points only.

The users of the system are assumed to be those both inside and outside the region of operation. It is also assumed that users accessing the system from outside the region of operation will mainly use PCs while inside, and will use the submitting, viewing and recommending function of sightseeing spot information, the submitting function for activity history, the support function for tour planning, and the navigation function in order to confirm routes. For users accessing the system from within the region of operation, it is assumed that they will be using mobile information terminals inside and outside, and that they will mainly use the submitting and viewing function of sightseeing spot information and the navigation function in order to visit sightseeing spots, which will both use the current location information of users. In addition, regarding the submitting function of sightseeing spot information, the use method of those with knowledge concerning the region of operation and those without will be different. It is assumed that the former, in addition to submitting sightseeing spot information that they already know about, will make submissions concerning the sightseeing spot evaluation using comment and tag functions. For the latter, it is assumed that they will visit sightseeing spots based on submitted information, and make submissions concerning the evaluation in the same manner as the former.

B. Operation tests and operation test evaluation

Before the operation directly via the web using PCs and mobile information terminals, a 1-week operation test was conducted with 5 students in their 20's, who belong to the authors' lab, as subjects. From the hearing survey results of the subjects after the operation test, one point of improvement was extracted, and the system was reconfigured regarding it only. Specifically, a breadcrumb list for each page was created to clearly show the location of users on the website. In addition, concerning the operation via smart glasses, a 1-day operation test using actually the smart glasses and the same test subjects in the region of operation was also conducted. From the hearing survey results of the subjects after the operation test, the only improvement was making the size of the letters displayed on the glass bigger.

C. Operation directly via the web using PCs and mobile information terminals

1) Operation overview and results

Firstly, the operation directly via the web using PCs and mobile information terminals was conducted. Whether inside or outside the region of operation, the operation of the system was advertised using the website of the authors' lab, and the tourism department of Kanagawa Prefecture and Yokohama City in addition to the Yokohama Convention and Visitors Bureau (Yokohama City Tourism Association) supported this study by distributing pamphlets and operating manuals. Users must register an account by entering their email address and password. If users use a registered account to access the system for the first time, their "nickname", "gender", "age" and "preferences" must be registered as user information. "Nicknames" do not have to be the user's real name, and by

allowing users to enter any half-width alphanumeric, users can remain anonymous.

Users will be moved to the homepage after registering user information, and functions including the submitting, viewing and recommending of sightseeing spot information, registering activity history, as well as navigation to single sightseeing spots will be made available. When activity history is registered, the support function of tour planning will be made available. Additionally, when tour plans are created, the navigation function of multiple sightseeing spot groups will be made possible based on the tour plan. When changing registered user information, by choosing the profile from the account, users will be moved to the update page of registered information. After 7 weeks of operation in normal mode assuming normal conditions, the operation in emergency mode assuming disasters was also conducted in the same region of operation with the same users for 1 week.

TABLE II shows the details of users during the 7-week operation as mentioned above, and Fig. 7 similarly shows the transition of the total number of users. The number of users gradually increased, and the total number of users were 86, with 40 male users and 46 female users. The percentage of users in their 20's was 67%, those in their 30% was 12 %, those in their 40's was 8%, and the total of those in their 20~40's occupied 87% of the total number of users. As shown in the 2015 Telecommunications (2015) [17], this is in harmony with the fact that the main users of general SNS are those in their 20~40's. After having each user use this system for a month, the evaluation by means of a web questionnaire survey was conducted.

2) Use of submitted information and the comment and tag functions

Fig. 7 also shows the transition of the number of submissions during the 7-week operation as mentioned above, and the significant increase in submissions from the 4th week was clear. This may because users who gradually got used to the system started submitting information they knew or thought was necessary halfway through the operation period. In addition, in order to solve the cold-start problem mentioned in Section IV, the 181 items of sightseeing spot information gathered by Ikeda et al. (2014) [15] was prepared as initial data. As the total number of submissions during the operation period was 170, a total of 351 sightseeing spot information items were accumulated in this system.

TABLE I. OPERATION PROCESS OF THE SYSTEM

Process	Aim	Period	Specific details
1. Survey of present conditions	To understand efforts related to tourism in the region for operation (Yokohama City)	December 2014 - March 2015	- Survey of government measures and internet services - Interviews with government departments responsible, tourist associations, etc.
2. System configuration	Configure the system in detail to suit the region for operation	April - July 2015	- Define system requirements - System configuration - Create operation system
3. Operation test	Conduct the system operation test	August 2015	- Create and distribute pamphlets and operating instructions - System operation test
4. Evaluation of operation test	Reconfigure the system based on results of interviews with operation test participants	August - September 2015	- Evaluation using interviews - System reconfiguration - Amendment of pamphlets and operating instructions
5. Operation	Carry out actual operation of the system	October - November 2015	- Appeal for use of the system - Distribution of pamphlets and operating instructions - System operation management
6. Evaluation	Evaluate the system based on the results of Web questionnaires, and the results of access analysis which used log data during the period of actual operation	November - December 2015	- Evaluation using Web questionnaires, access analysis which used log data - Identification of measures for using the system even more effectively

TABLE II. OUTLINE OF USERS AND RESPONDENTS TO THE WEB QUESTIONNAIRE (OPERATION DIRECTLY VIA THE WEB USING PCS AND MOBILE INFORMATION TERMINALS)

	Aged 10 to 19	Twenties	Thirties	Forties	Fifties	Sixties and above	Total
Number of users (people)	4	59	7	10	2	4	86
Number of Web questionnaire respondents (people)	3	37	2	6	1	2	51
Valid response rate (%)	75.0	62.7	28.6	60.0	50.0	50.0	59.3

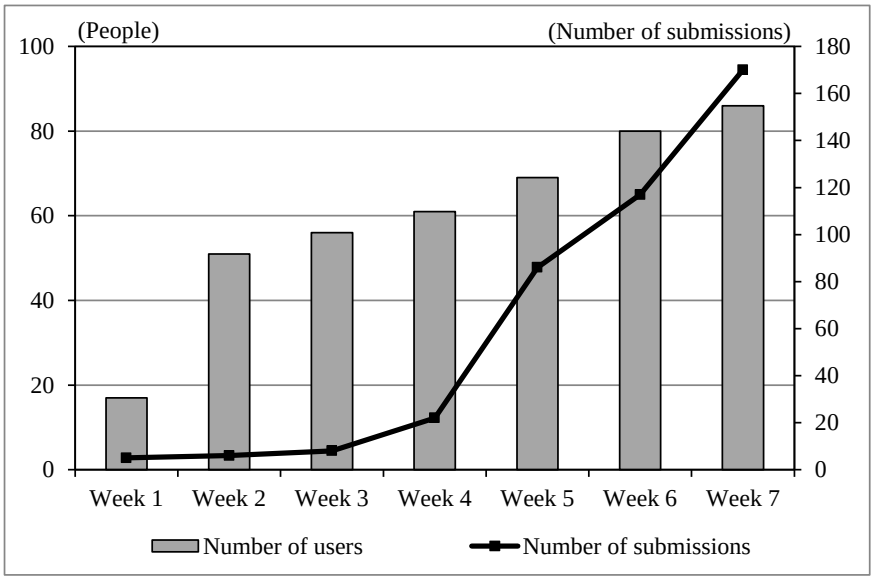


Fig. 7. Changes in the number of users and number of submissions during the operation period (operation directly via the web using PCs and mobile information terminals)

TABLE III. SUBMISSIONS OF INFORMATION, CLASSIFIED BY TOURIST SPOT CATEGORY (OPERATION DIRECTLY VIA THE WEB USING PCS AND MOBILE INFORMATION TERMINALS)

Category	All submissions		Submission during operation		Initial data	
	Number	Percentage (%)	Number	Percentage (%)	Number	Percentage (%)
Food and drinks	126	35.9	78	45.9	48	26.5
Shops	62	17.7	36	21.2	26	14.4
Entertainment	25	7.1	21	12.4	4	2.7
Event	9	2.6	5	2.3	4	2.7
Landscape	19	5.4	4	2.4	15	7.3
Art	84	23.9	21	12.4	63	34.8
Recreation	26	7.4	5	3.5	21	11.6
Total	351	100.0	170	100.0	181	100.0

TABLE IV. OUTLINE OF USERS AND RESPONDENTS TO THE WEB QUESTIONNAIRE (OPERATION VIA THE SMART GLASSES)

	Aged 10 to 19	Twenties	Thirties	Forties	Fifties	Sixties and above	Total
Number of users (people)	4	10	3	8	7	2	34

TABLE III shows submitted information concerning sightseeing spots by category. As shown in TABLE III, during the operation period, although many submissions were of categories including food and drinks (78 items, 46%) and shops (36 items, 21%), information concerning all categories were submitted. Additionally, with almost all submitted information, related images were also submitted. From these results, it can be said that various sightseeing spot information is submitted and information attached with images for the recommendation of sightseeing spots according to the preference of each user is accumulated in line with the purpose of this system.

11 comments were made and 111 tags were registered to sightseeing spot information by users during the operation period. Although the comment function is not used often, from the fact that evaluation using the tag functions was made often, it is understood that communication between users is made mainly through the tag functions. This may because the tag function is easier to use as users only need click on a tag to give the submission more weight, while the comment function requires users to enter sentences.

D. Operation via smart glasses

Between Yamashita Park and the Yokohama Red Brick Warehouse located within the region of operation, on December 11th and 18th, 2015, the operation via the smart glasses was conducted with tourists as subjects. Users put the smart glasses on and received navigation for 600m between the above-mentioned two places. The reason why such a route was chosen is because there are no cars which enables users to receive navigation safely. Additionally, in consideration for the safety of users, an escort was assigned to all users.

TABLE IV shows the details of users during the 2 days of the operation period as mentioned above, and the total number of users was 34, with 18 male users and 16 female users. When divided according to age, although those in their 20's were the

most numerous occupying 29% of the total number of users, the age of users were scattered, and no one had experience using the smart glasses. Just after the operation, all users were required to answer the web questionnaire survey.

VII. EVALUATION

In this section, based on the questionnaire survey results as shown in the overview in TABLEs II and IV, the evaluation concerning the system using the web system and the smart glasses is conducted. Next, based on the access analysis results using the log data from during the operation, the evaluation concerning the activity support during both normal conditions and disasters is conducted. In addition, based on these evaluation results, points of improvement for this system are extracted.

A. Evaluation based on the questionnaire survey results concerning the operation directly via the web using PCs and mobile information terminals

1) Evaluation concerning the use of the system

Fig. 8 shows the evaluation results concerning the use of this system in normal conditions and disasters. Regarding the usefulness in tourist spots in normal conditions, all answered were either "I agree" and "I somewhat agree", and 74% answered "I agree" which is a significantly high number. When switching from normal mode assuming normal conditions to emergency mode assuming disasters, although a high number of 88% answered "I agree" or "I somewhat agree" concerning the smooth use, 12% answered "I somewhat disagree" or "I completely disagree". Because the functions of this system used in the event of disasters completely differ from those used in normal conditions. However, concerning the usefulness in the event of disasters in tourist spots, with all answers being either "I agree" and "I somewhat agree", a significantly high percentage of 69% answered "I agree". From the information above, regarding the support of both sightseeing in normal conditions and the evacuation in the event of disasters, it can be said that this system is effective. Also, concerning whether users would like to use this system in the future, as 96% answered "I agree" or "I somewhat agree", the continuation of this system's operation in the future can be expected.

2) Evaluation concerning the function of the system

a) Evaluation of use frequency according to the function

Fig. 9 shows the evaluation results concerning the frequency of use according to the function while in normal mode assuming normal conditions, and also shows the aforementioned according to the type of information terminal mainly used and those used in this entire system. In particular, the ratio of those who chose the top 2 frequently used functions will be shown according to the function. This entire system has a high percentage of 53% in the "viewing function of sightseeing spot information" followed by the "recommendation function of sightseeing spot information" with 45%, and this shows that the system's main purpose of use is the gathering of sightseeing spot information. In addition, regarding the use tendency of each function according to the type of information terminal, the results for PCs were the "viewing function of sightseeing spot information (53%)", the "recommendation function of sightseeing spot information (50%)" and the "support function of tour planning (39%)". For mobile information terminals, although the percentage of the "viewing function of sightseeing spot information (53%)" is the same as that of PCs, it is followed by the "navigation function (40%)", the "recommendation function of sightseeing spot information (33%)" and the "support function of tour planning (33%)". Excluding the basic function related to the submitting and viewing of sightseeing spot information, as the support of tour planning and the recommendation function of sightseeing spot information for PCs used indoors, and the navigation function for mobile information terminals used indoors and outdoors are designed to be the main functions of this system, the results mentioned above show that this system is used as planned when the system was being designed (see Section IV).

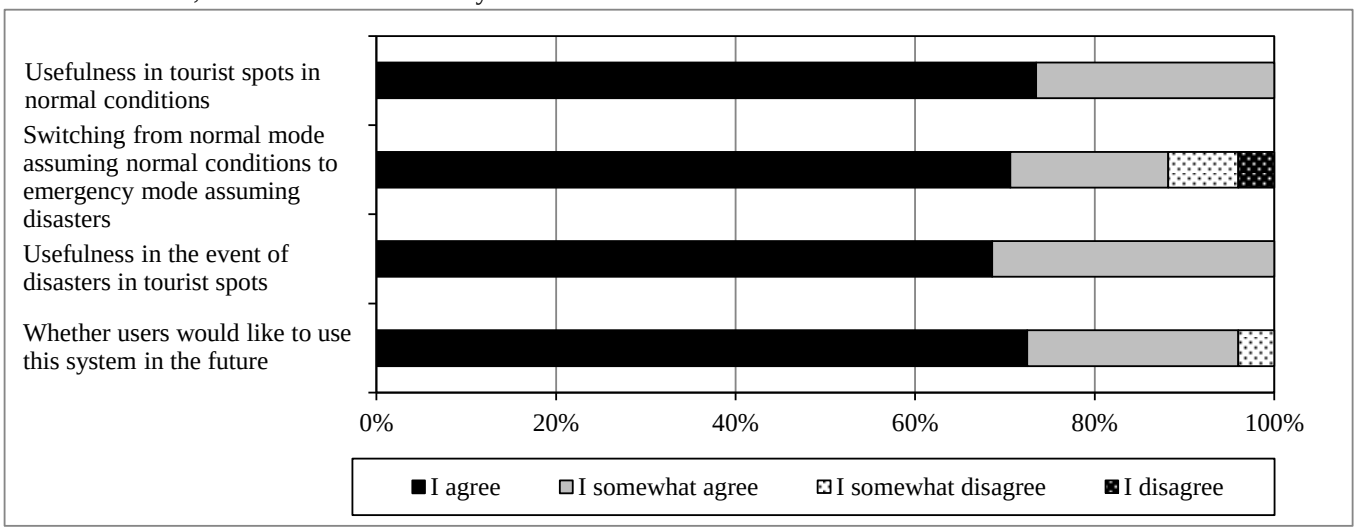


Fig. 8. Evaluation results concerning the system use

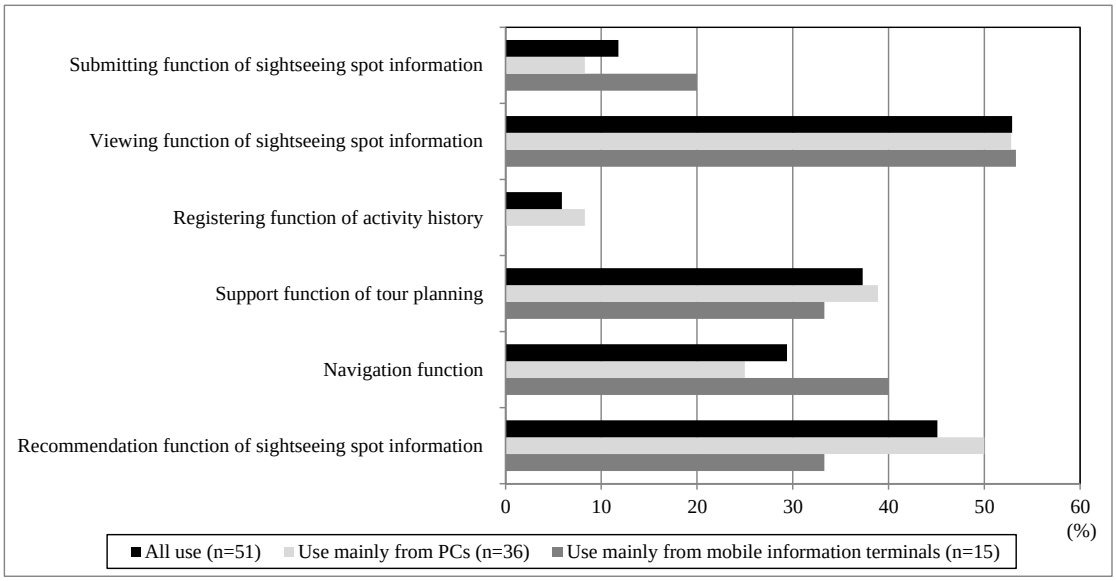


Fig. 9. Evaluation results concerning the use frequency according to the function while in normal mode assuming normal conditions

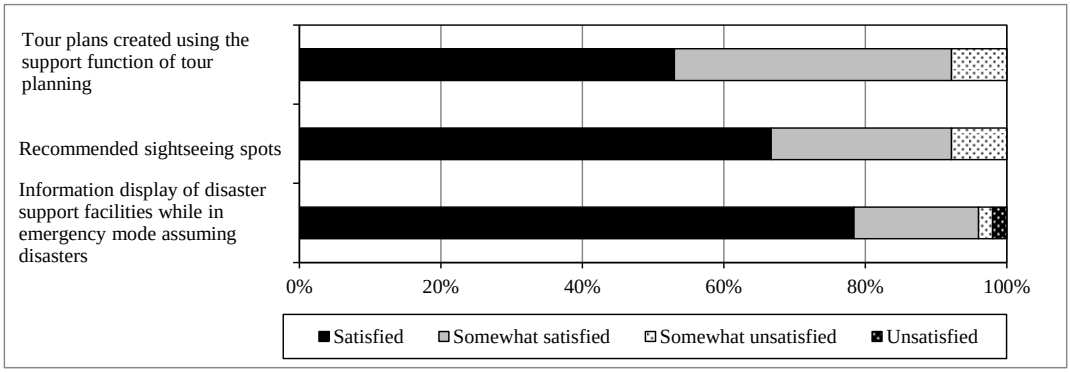


Fig. 10. Evaluation results concerning the satisfaction level of the original features

b) Evaluation concerning the satisfaction rate of the original function of this system

Fig. 10 shows the evaluation results only of the results provided to users, among the original functions of this system, in which the satisfaction rate can be questioned (the support function of tour planning and the recommendation function of sightseeing spot information used while in normal mode assuming normal conditions, in addition to the information display function of disaster support facilities while in emergency mode assuming disasters). Concerning the tour plans created using the support function of tour planning and the recommended sightseeing spots, 93% answered "satisfied" or "somewhat satisfied" for both, and the result of the latter was especially good with 67% answering "satisfied". From these results, it can be said that the recommendation system integrated into this system has provided appropriate sightseeing spot information for users. Additionally, regarding information display of disaster support facilities while in emergency mode assuming disasters also, as 96% answered "satisfied" or "somewhat satisfied", it can be assumed that this system can provide users with disaster support facility information appropriately in the event of disasters.

B. Evaluation based on the questionnaire survey results of the operation via smart glasses

1) Evaluation concerning the use of the system

Fig. 11 shows the evaluation results concerning the system using the smart glasses. 94% answered "easy" or "relatively easy" regarding the usability of the smart glasses, and from the fact that all users had no experience using the smart glasses previously, it can be said that the use of the smart glasses concerning this system is easy even for users who have never used it before. Concerning the suitability of smart glasses in comparison to smartphones while sightseeing, in addition to the suitability of navigation by means of the smart glasses, 91% answered "suitable" or "relatively suitable" for both situations. The former had especially high results as 74% answered "suitable". The reason for this is that, in addition to the usability of the smart glasses as mentioned above, as the smart glasses, unlike mobile information terminals, can display information right in front of users, users can take in information while looking ahead instead of looking down. Therefore, in order to realize efficient and safe navigation which is the aim of this study, it is beneficial to use the smart glasses for this system instead of only using mobile information terminals.

2) Evaluation concerning the safety of the smart glasses

Fig. 12 shows the evaluation results concerning the safety of the smart glasses. Regarding whether the information display on the smart glasses obstructs the view of users, although 79% answered "Not obstructed" or "Not obstructed greatly", the other 21% answered "somewhat obstructed". On the other hand, regarding whether users felt any danger while walking with the smart glasses on, 91% answered they "did not feel any danger" or "did not feel any great danger". From these results, it can be assumed that although the users' view was somewhat obstructed by the navigation information displayed on the bottom half of the screen, it did not obstruct the view in a way that would make users' feel endangered. Therefore, concerning this system, it can be said that safe navigation using the smart glasses was realized.

C. Evaluation concerning the activity support for users

1) Overview of access analysis

In this study, by conducting an access analysis using log data gathered during the operation directly via the web using PCs and mobile information terminals, the evaluation focusing on the number of times accessed as well as the access method will be made. This study will incorporate the API of Google Analytics to the developed program, and then the access analysis will be conducted. Google Analytics is a web access analysis service provided by Google, and it can obtain access information of users concerning the website. In order to use Google Analytics, concerning the website used in the access analysis, a tracking code must be added directly to the HTML of each page.

2) Evaluation based on the access analysis results

First, the access log analysis of users concerning operation while in normal mode assuming normal conditions was conducted. The total number of sessions was 358, and concerning the information used as the method for accessing this system, PCs were 77% and mobile information terminals were 23%. TABLE V shows the number of times each function was accessed, and the most accessed was the "viewing function of sightseeing spot information (27%), " followed by the "support function of tour planning (20%)" and the "register function of activity history (19%)". From these results, it can be said that PCs are used as the main access method for those outside the region of operation, and that the gathering of information and tour planning are the main purpose of use. This is in line with the assumption of users' use method outside the region of operation, as mentioned in Section VI.

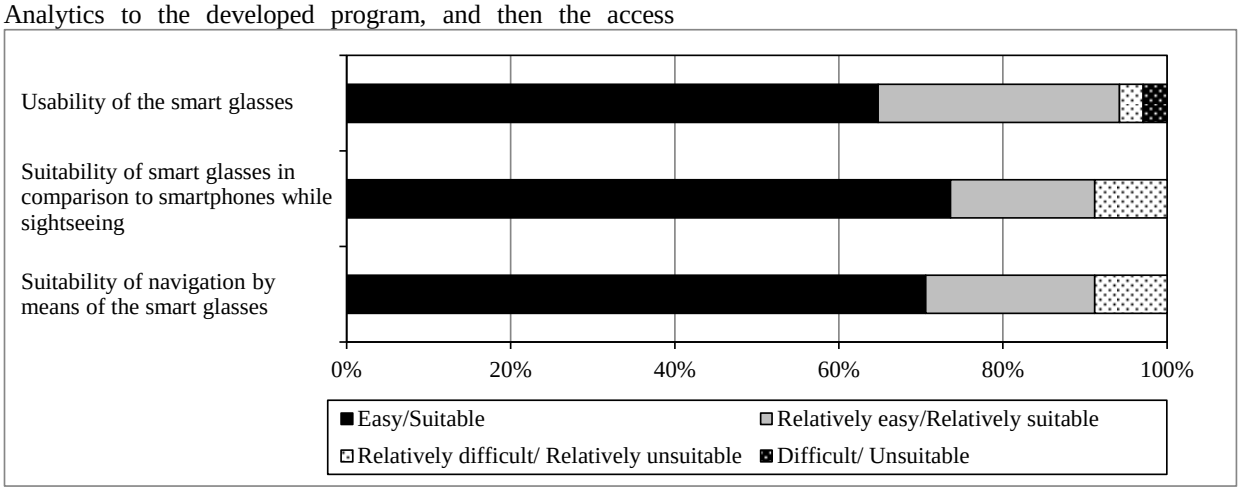


Fig. 11. Evaluation results concerning the use of the system using smart glasses

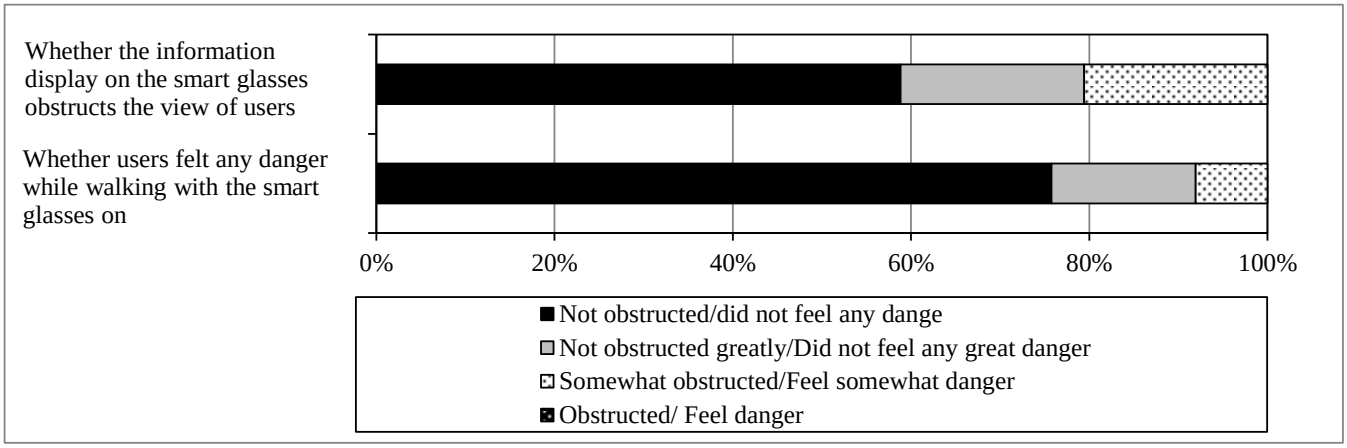


Fig. 12. Evaluation results concerning the safety of the smart glasses

TABLE V. TOP 10 OF THE MOST VISITED PAGES ACCORDING TO THE TYPE OF INFORMATION TERMINALS

Functions for normal mode assuming normal conditions	Total number of accesses (556)
Submitting function of sightseeing spot information	13.8
Viewing function of sightseeing spot information	27.0
Registering function of activity history	18.9
Support function of tour planning	19.6
Navigation function	7.3
Recommendation function of sightseeing spot information	13.4
Functions in emergency mode assuming disasters	Total number of accesses (116)
Viewing function of support facilities in the event of disasters	44.5
Search function of support facilities near users in the event of disasters	22.2
Navigation function to the closest evacuation sites	33.3

Next, the access log analysis of users concerning the operation while in emergency mode assuming disasters was conducted. The total number of sessions during the 1-week operation period was 28, and concerning the information terminals used as a method of access to this system, the ratio of PCs and mobile information terminals were 5:5. From the increase in percentage of mobile information terminals in comparison to normal conditions, it can be said that the tendency for this system's users to use mobile information terminals became stronger when disasters are assumed. The reason for this could be that users, assuming disasters in tourist areas, thought to use this system mainly by means of mobile information terminals. Concerning the number of times each function was accessed, although the "viewing function of disaster support facilities" was accessed by almost half with 45%, as functions using location information including the "search of disaster support facilities near users" and the "navigation function to the closest evacuation site" are 22% and 33%, it can be said that functions requiring location information were also used actively.

D. Extracting solutions

From the evaluation results in this section, the two points of improvement for this system can be summarized as shown below.

(1) Implementation of the automatic switch function to emergency mode

By obtaining information of the disasters in the region of operation and reflecting this in the system, the automatic switch from normal mode to emergency mode can be made possible. This will not only lighten the load of the system manager, but will enable the switch to emergency mode and support the evacuation of users, regardless of the manager's situation in the event of disasters. However, considering the fact that the system may be switched to emergency mode because of incorrect information, the implementation of a function that allows the manager to manually switch it back is also necessary.

(2) Color classification of displayed routes concerning navigation

When navigating multiple sightseeing spots on the Web-GIS, by color-coding each displayed route between sightseeing spots, the discernment of routes are made easier. This enables the operability of the navigation function to improve, and a more efficient sightseeing support is also made possible.

VIII. CONCLUSION

The conclusion of this study can be summarized into three points as shown below.

(1) In order to support sightseeing in normal conditions and evacuation in the event of disasters by integrating SNS, Twitter, Web-GIS, the recommendation system and the smart glasses, as well as gathering high real time information concerning urban sightseeing spots, a navigation system that can actively change routes was designed and developed. By means of this, concerning both normal conditions and disasters, the accumulating, sharing, recommending of information in addition to navigating users to their destination were made possible. In addition, the center part of Yokohama City in Kanagawa Prefecture was chosen as the region of operation, and details of the system were organized after conducting a survey of the current situation.

(2) Because the operation directly via the web using PCs and mobile information terminals was conducted over a period of 8 weeks, a 1-week operation test was conducted beforehand, and the system was reconfigured based on the extracted points of improvement. It was assumed that all users were over 18 regardless of whether they were located inside or outside the region of operation, and among the 86 users, a total of 87% were in their 20-40's, and the total number of submitted information was 170. Additionally, the operation via smart glasses was conducted over the course of 2 days, and concerning the 34 users who participated, many were from different age groups and all users had no experience using the smart glasses.

(3) From the results of the Web questionnaire survey to users after the operation, it was clear that this system can appropriately support both sightseeing in normal conditions as well as evacuation in the event of disasters, and the safe and efficient navigation using the smart glasses has been realized. In addition, from the results of the access analysis using the log data form during the operation, it was shown that users, especially those outside of the region of operation used this system as assumed in normal conditions. During the operation assuming disasters, it was also shown that the tendency for users to use mobile information terminals to access this system was stronger and functions which require location information as well as the viewing function of information were actively used.

For future research issues, the new implementation of functions that support sightseeing in a more efficient way as suggested in Section VII, the increase in achievements by operating this system in other urban sightseeing spots, and the improvement in the significance of use can be raised.

ACKNOWLEDGMENT

In the operation of the dynamic real-time navigation system and the web questionnaires of this study, enormous cooperation was received from those mainly in the Kanto region such as Kanagawa Prefecture and Tokyo Metropolis. We would like to take this opportunity to gratefully acknowledge them.

REFERENCES

[1] Y. Kurata, "Introducing a hot-start mechanism to a Web-based tour planner CT-Planner and Increasing its coverage areas", Papers and Proceedings of the Geographic Information Systems Association of Japan, Vol.21, CD-ROM, 2012.

[2] J. Sasaki, T. Uetake, M. Horikawa and M. Sugawara, "Development of personal sightseeing support system during long-term stay", Proceedings of 75th National Convention of IPSJ, pp.727-728, 2013.

[3] T. Fujitsuka, T. Harada, H. Sato and K. Takadama, "Recommendation system for sightseeing plan using pattern mining to evaluate time series action", Proceedings of the Annual Conference on Society of Instrument and Control Enginnering 2014, SS12-10, pp.802-807, 2014.

[4] T. Ueda, R. Ooka, K. Kumano, H. Tarumi, T. Hayashi and M. Yaegashi, "Sightseeing support system to support generation / sharing of sightseeing information", The Special Interest Group Technical Reports of IPSJ: Information system and Social environment (IS), 2015-IS-131(4), pp.1-7, 2015.

[5] J. M. Noguera, M. J. Barranco, R. J. Segura, and L. Martinez, "A mobile 3D-GIS hybrid recommender system for tourism", Information Sciences, Vol.215, pp.37–52, 2012.

[6] M. Ye, P. Yin, W. C. Lee and D. L. Lee, "Exploiting geographical influence for collaborative point-of-interest recommendation, Proceedings of the 34th international ACM SIGIR conference on Research and Development in Information Retrieval, pp. 325-334, 2011.

[7] Q. Yuan, G. Cong, Z. Ma, A. Sun and N. M. Thalmann, "Time-aware point-of-interest recommendation, Proceedings of the 36th International ACM SIGIR Conference on Research and Development in Information Retrieval, pp.363-372, 2013.

[8] M. Chen, F. Li, G. Yu and D. Yang, "Extreme learning machine based point-of-interest recommendation in location-based social networks, Proceedings of ELM-2015, Vol. 2, pp. 249-261, 2016.

[9] T. Yanagisawa and K. Yamamoto, "Study on information sharing GIS to accumulate local knowledge in local communities", Theory and Applications of GIS, Vol.20, No.1, pp.61-70, 2012.

[10] H. Nakahara, T. Yanagisawa and K. Yamamoto, "Study on a Web-GIS to support the communication of regional knowledge in regional communities: Focusing on regional residents' experiential knowledge", Socio-Informatics, Vol.1, No.2, pp.77-92, 2012.

[11] S. Yamada and K. Yamamoto, "Development of Social Media GIS for information exchange between regions", International Journal of Advanced Computer Science and Applications, Vol.4, No.8, pp.62-73, 2013.

[12] T. Okuma and K. Yamamoto, "Study on a Social Media GIS to accumulate urban Disaster Information: Accumulation of Disaster Information during normal times for disaster reduction measures", Socio-Informatics, Vol.2, No.2, pp.49-65, 2013.

[13] T. Murakoshi and K. Yamamoto, "Study on a Social Media GIS to support the utilization of disaster information : For disaster reduction measures from normal times to disaster outbreak times", Socio-Informatics, Vol.3, No.1, pp.17-30, 2014.

[14] K. Yamamoto and S. Fujita, "Development of Social Media GIS to support information utilization from normal times to disaster outbreak times", International Journal of Advanced Computer Science and Applications, Vol.6, No.9, pp.1-14, 2015.

[15] T. Ikeda and K. Yamamoto, "Development of Social Recommendation GIS for tourist spots", International Journal of Advanced Computer Science and Applications, Vol.5, No.12, pp.8-21, 2014.

[16] D. Jannach, M. Zanker, A. Felfernig and G. Friedrich, "Recommender systems: An introduction", Cambridge University Press, U.K., 2011.

[17] Ministry of Internal Affairs and Communications of Japan, "2015 White paper -Information and communications in Japan", Tokyo, 2015.

MIMC: Middleware for Identifying & Mitigating Congestion Level in Hybrid Mobile Adhoc Network

P. G. Sunitha Hiremath
Assc Prof
Dept. of Information Science & Engg., BVBCET, Hubli, India

C.V. Guru Rao
Director of Evaluation,
S.R. Engg. College, Warangal, India

***Abstract*—Adoption of middleware system to solve the congestion problem in mobile ad-hoc network is few to find in the existing system. Research gap is found as existing congestion control mechanism in MANET doesn't use middleware design and existing middleware system were never investigated for its applicability in congestion control over the mobile ad-hoc network. Therefore, we introduce a novel middleware system called as MIMC or Middleware for Identifying and Mitigating Congestion in Hybrid Mobile Adhoc Network. MIMC is also equipped with novel traffic modeling using rule-based control matrix that not only provides a better scenario of congestion but also assists in decision making for routing, which the existing techniques fails. This paper discusses the algorithms, result discussion on multiple scenarios to show MIMC perform better congestion control as compared to existing techniques.**

Keyword*—Middleware; Congestion Control; Traffic Management; Hybrid Mobile Adhoc network

I. INTRODUCTION

The mobile ad-hoc network has played a huge contribution in the area of the wireless ad-hoc network, and it has been a point of major research area since last decades in wireless networking. Till date, there has been researching on multiple problems in mobile ad-hoc network e.g. routing issues [1][2], energy issues [2], security issues [3], load balancing issues [4], congestion control issues [5], etc. Out of all the problems, routing and security have attained much attention. We will like to discuss the congestion problems in the mobile ad-hoc network that occurs due to superfluous number of data bigger than channel capacity. An adverse effect of congestion is a loss of data packets, intermittent links, interference, etc. [6]. Hence, it is highly essential to perform controlling of congestion in a dynamic topology of the mobile ad-hoc network. One of the biggest problems with the existing congestion control protocol is non-consideration of routing schema [7]. A closer look at the trends of research shows that majority of the studies towards congestion control is focused on homogeneous network and very less focus is laid on to a heterogeneous network. This evidence can also be found by observing less number of work being carried out in designing middleware system in the mobile ad-hoc network. The evolution of middleware is not new, but they are more involved in the theoretical study and less in practical implementation. Problems of middleware to be more focused on resource management is seen in theory and very less on implementation papers. A middleware must be able to configure multiple forms of resources in many ways [8][9]. However, usage of middleware is not that much clear in the research area of the mobile ad-hoc network. Till date, there is no discussion of any middleware systems over a mobile ad-hoc network that can perform identification of the state of traffic congestion and provides a solution to mitigate it. Hence, this paper presents a novel middleware system that can perform identification of congestion in the mobile ad-hoc network. The design principle of the proposed middleware system is meant to achieve i) seamless and cost-effective monitoring of traffic condition for mobile ad-hoc network and its applications (e.g. vehicular network), ii) assists in relaying information about the congestion level of traffic in the entire network, iii) less overhead in message dissemination, etc. The paper presents a cost-effective technique of middleware services which provides interoperability, energy efficiency, as well as congestion control mechanism in the hybrid mobile ad-hoc network. Section II discusses the background of the study where the discussion of the recent system of traffic congestion control in the mobile ad-hoc network is carried out followed by brief discussion of problems in Section III. Section IV discusses the contribution of the proposed system regarding adopted research methodology. Section V discusses the algorithm discussion followed by result discussion in Section VI. Finally, the findings of the proposed study are summarized in Section VII.

II. RELATED WORK

This section discusses about various significant studies conducted towards normalizing the traffic behavior in hybrid mobile adhoc network. The primary motive of this section is to showcase the research papers towards addressing the problems of traffic management using middleware-based solution for hybrid mobile adhoc network.

Vadivel and Bhaskaran [10] has presented a recent technique of formulating congestion-free routes as well as to the study has also introduced a reliable routing in the mobile ad-hoc network. The author has used a mechanism of disjoint path construction for balancing the traffic load. The study outcome was found to posses better communication performance. Study towards addressing congestion problem was also discussed by Sirajuddin et al. [11], where the authors has used a TCP-based congestion control management over a mobile ad-hoc network. The technique mainly forms a network followed by propagation of congestion status and finally by route maintenance depending on the status of traffic load. The study outcome was compared with conventional AODV on packet delivery ratio and routing overhead.

Pashchenko et al. [12] has presented a model that is designed for cloud model for networking. However, the entire modelling was carried out using the concept of middleware in mobile ad-hoc network. The discussion laid by the author is highly helpful as all the discussed formal models in this research papers can be readily used in any form of wireless network for traffic management including QoS as well as security. The study carried out by Chen et al. [13] has introduced a cross-layered architecture incorporated on the integrated scheme of congestion management and QoS scheduling in a wireless network which supports multihop. The technique uses Differentiated Queuing Service along with partial TCP services for this purpose. The study outcome was testified for reduced delay and increased delivery ratio.

Sreenivas et al. [14] has presented an approach that can perform controlling of traffic congestion over the mobile ad-hoc network after enhancing TCP. According to this technique, the significant information of the network status is identified by destination node that further transmits to the source node in the form of response. The study outcome was evaluated to posses better throughput and delay performance. Greco et al. [15] has presented a technique to overcome the problem of latency considering the case study of multimedia streaming over the mobile ad-hoc network. The technique uses a cross layer-based approach where coding of multimedia is carried over MAC layer followed by optimization over the state of congestion in the mobile ad-hoc network. The author has developed a distortion model powered by sophisticated mathematical modelling using graph theory to perform congestion identification and mitigation. The study outcome was evaluated on relative frequency over increasing delay and probability density over PSNR. Bhaduria and Sharma [16] has presented a framework that can control congestion in a mobile ad-hoc network with the aid of agent-based services. The technique allows selection of less-congested route

Qin et al. [17] has developed a middleware approach for catering up the need of the heterogeneous networking environment, which is based on the self-intelligence building process to accomplish better communication performance. The system uses observe-analyze-adapt methodology for developing the formal node model. Study towards the evolution of middleware was seen in the work of Kamisinski et al. [18]. The technique was meant for sensor nodes for processing complex information entrapped by a sensor using distributed middleware system. Testified over multiple mobility environments, the technique allows dynamic selection of routing algorithm depending on the need of applications. Another study of middle concerning about traffic management was carried out by Denker et al. [19]. The study mainly focuses on cyber-physical systems and highlights about its dependencies. Pease [20] has developed a middleware framework called as ROAM which intends to support energy communication using cross layered approach in mobile ad-hoc network.

The study uses cross-layer approach to identify the new route and performs blacklisting of the path which is inflicted with channel fading problem. Another middleware approach was presented by Lopez et al. [21] where the author has presented a technique to maintains the portability of the routing protocols in mobile adhoc network. The technique has used multicast protocol design to develop a group communication system in mobile adhoc network. The study outcome was evaluated on a real test bed on some the message being used for transmission. Liu [22] has presented a study of middleware system for a wireless network using context-based factors. The prime purpose of this system is to present a technique with better supportability of distributed synchronization and dynamic reconfiguration.

Hence, it can be seen that there are various studies car-ried out for congestion control as well as the various design of middleware to upgrade the communication performance of mobile adhoc network. Each study has its beneficial factor regarding communication performance and problems being discussed in the research papers. However, these studies are also associated with certain loopholes when it comes to congestion control over hybrid the mobile ad-hoc network using middleware-based approach. The next section highlights significant points of limitations and justifies the evolution of problem statement.

III. Problem Description

This section discusses the problems being identified after reviewing the existing system discussed in prior section.

A. Less Applicability of existing Congestion Control:

Although, there is a certain level of work being carried out addressing congestion issues over heterogeneous mobile ad-hoc network e.g. [23], none of them has addressed the interoperable part of it. Hence, heterogeneous network without the inclusion of interoperability is not the purely heterogeneous network. Hence, such algorithms are less likely to be effective when implemented over the real-world environment. The existing techniques of congestion control are more inclined towards homogeneous network and very less towards heterogeneous networks.

B. Frequent usage of Cross Layer Approach:

It is highly essential that the design approach of a middleware must has the knowledge of the topology to use for a multihop communication system, and it maintains its efficiency. Hence, adoption of cross-layer approach assists over kernel space in communicating with networking protocols. This approaches, therefore, violates the stringent layering mechanism of the network stack in the mobile ad-hoc network that obstructs the true performance of middleware services.

C. Lew focus on hybrid networks:

The majority of the study towards traffic management, congestion control, load balancing task scheduling, etc. are either focused on reactive or proactive routing protocols in the mobile ad-hoc network. However, similar area of research on hybrid routing methodology was less considered. There exists certain few research works focusing on hybrid routing e.g. [24][25], but they are fewer standards in nature or has a very limited scope of future enhancement. Hence, study on congestion control is quite a few to find on hybrid networks.

A closer look at the problem identification points says that although there has been work carried out for congestion control in mobile ad-hoc network, there is a bigger trade-off between congestion control and adoption of the hybrid protocol as well as middleware design. Middleware design can act as a robust

bridge of interoperability and thereby can act upon heterogeneous mobile ad-hoc network. However, in the past, there is almost no research manuscript being found for addressing the problem of congestion using middleware-based approach. It is going to be completely a new arena of research, whereby hybrid network adoption will be the first to be tested for congestion control using middleware design. Therefore, the problem statement of the proposed system can be represented as –"*It is a novel and challenging task to develop a modelling of middleware system that can carry out congestion control mechanism over hybrid network to attain better communication performance.*"

IV. Proposed MIMC Methodology

The proposed study is an extension of our prior studies where we has introduced two different characteristics on our middleware design i.e. interoperability and energy efficiency. Our first model MERAM [26] is focused on incorporating interoperability in its middleware design whereas the second model MEEM [27] is more focused on developing energy efficient features. The proposed model MIMC aims at incorporating congestion identification and mitigation characteristics for hybrid mobile ad-hoc network.

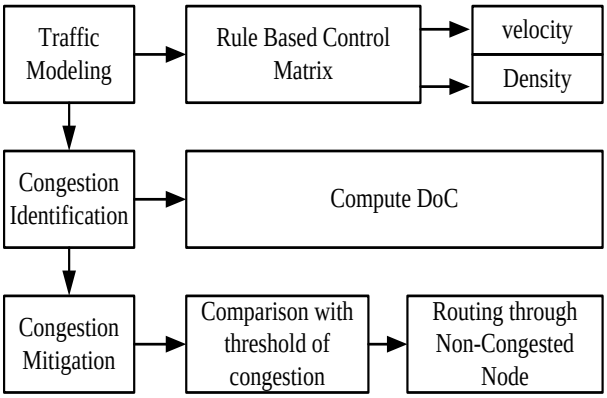


Fig. 1. Proposed Schema of MIMC

The proposed study is developed using analytical research methodology where the mechanism of MIMC is divided into three core modules i.e. traffic modelling, congestion identification, and congestion mitigation. Traffic modelling is carried out using multi-valued logic which provides the better shape of hybrid mobile ad-hoc network on real-world applications using mobile nodes. The system uses pre-defined ruleset called as Rule Based Control Matrix which gives inferences based on velocity and density of node. The second module is about exploring the presence of congestion in neighbor nodes. MIMC doesn't attempt to spend its resources exploring the source of congestion rather it uses its resources just to find the alternative routes. Hence, it is much cost effective and provides a better response time for exploring non-congested routes using a new term called as Degree of Congestion or DoC. The last step is to perform mitigation of congestion, which will mean filtering the message that estimates the congestion and compares it with a threshold. An elaborated information about its implementation is discussed in next section.

V. Algorithm Implementation

This section discusses the algorithm design and implementation that was used in proposed middleware system in the hybrid mobile ad-hoc network. It should be noted that the proposed middleware system is meant to carry out following operation i.e. i) traffic modelling, ii) congestion identification and iii) congestion control. Following are the detailed information about it.

a) Traffic Modelling: - MIMC make use of node-to-node communication system to perform traffic modelling along with communication. The algorithm to develop traffic modelling takes the input of v (velocity), p (position), and t (time interval), which after processing yields simulated a version of the traffic in hybrid mobile ad-hoc network. For this purpose, we develop a simple control message msg which will carry velocity (v) and positional (p) information of source node and should keep on exchanging with its neighbor nodes after a specific interval of time t.

Algorithm for Traffic Modelling

Input: v (velocity), p (position), t (time interval), d (density of nodes), *RBCM* (Rule Based Control Matrix)

Output: *traff* (traffic)

Start

1. Init v, p, t
2. Exchange beacons after every t-sec

$$i(msg[v,p]) \Leftrightarrow j(msg[v,p])$$

3. $v = \sum_{x=1}^{4} v_x \;\&\; d = \sum_{y=1}^{4} d_y$
4. f(RBCM)←[v, d]
5. traff→$simu_{scope}[f(RBCM)]$

End

Fig. 2. Algorithm for Traffic Modelling

To overcome the problem of topological dynamicity of the mobile nodes, we will use a novel rule-based mechanism based on simple mathematical logic to carry out traffic modelling. Our traffic modeling is not the simply distribution of nodes with mobility, but we emphasize more on the distinctive behavior of a node by monitoring the types of beacons msg being relayed by the nodes. For this purpose, we develop a simple matrix called as RBCM or Rule Based Control Matrix which discretely defines the various logical inference of the ruleset for velocity (v) and density (d). The algorithm exchanges of the control message among the communicating node (Line-2), which allows the mobile nodes to get updates status of the congestion level. The system considers 4 different states of velocity (fast, normal, slow, and very slow) whereas there are four discrete states for node density (i.e. less, medium, high, and very high) (Line-3).

TABLE I. FORMATION OF RULE-BASED CONTROL MATRIX

Density		Degree of Congestion (DoC)			
		Velocity of mobile node			
		Fast	Normal	Slow	Very Slow
	Less	Free	Free	Free	Slight
	Medium	Free	Slight	Slight	Moderate
	High	Free	Slight	Moderate	Moderate
	Very High	Slight	Moderate	Moderate	Severe

A function is developed for RBCM (Rule-Based Control Matrix) whose matrix formation is tabulated in Table.1, which creates a simple inference rules based on the types of velocity and node density as input arguments. The algorithm computes velocity from a distance traveled at a specific interval of time. Computation of node density is quite a difficult task. Depending on the number of the controlled messages received, each mobile node computes the total number of neighbor nodes at that period as the node density. The mobile nodes estimate the velocity factor by dividing instantaneous velocity with velocity limit assigned for the path. Hence, it is possible to calculate the highest number of mobile nodes in every path using a range of transmission. Hence, for a given section of a route, every mobile node can calculate the number of mobile nodes in that particular route segment by the ratio of total existing neighbor nodes in route section with the highest number of mobile nodes that can be accommodated in that particular route. The advantage of this traffic modelling is its simple inference mechanism that allows more flexibility of investigating middleware system for hybrid mobile ad-hoc network.

b) Congestion Identification: - The proposed MIMC adopts node-to-node interaction system to find the Degree of Congestion (DoC). The prior traffic model allows confirming this degree of congestion using RBCM. Therefore, the control message used msg can also be termed as a message for identifying (or estimating) congestion degree. The proposed mechanism initially finds the size of the control message msg with predefined threshold used for congestion i.e. Th (Line-1). The algorithm computes Degree of Congestion using neighboring nodes (NN) and frequency intervals (FI) (Line-2). The mobile node that computes DoC is considered to be mobile reference node and is used for computation of DoC. If the Degree of Congestion is found to be more than the threshold (Line-3) than it represents an event of congestion, however, it doesn't give a clear view of its location. Moreover, MIMC do not involve route acknowledgment messages to state confirmation of the message delivery. Further, overhead is minimized by lowering the frequency (Line-5). The algorithm than continues its search for all the nodes whose Degree of Congestion level is more than the threshold to identify the location of those nodes. Only the nodes with a higher value of DoC will be considered for performing retransmission process. However, if any mobile nodes pass within the transmission range of this node than it halts its transmission (Line-6-9).

Algorithm for Congestion Identification
Input: DoC (Degree of Congestion), F_I (frequency intervals), N_N (Number of neighbor nodes), E_D (Effective Distance), T_r (Transmission range), W_{node} (wavelength of node), f (frequency), Th (Threshold)
Output: Congestion identification
Start
1. msg → size(msg)
2. DoC=[F_I, N_N]
3. if DoC>Th
4. Congestion Identified
5. f=(f-0.1)
6. If node(DoC>Th))
7. node→retrans
8. else
9. Permit routing
10. compute relay distance of nodes

$$E_D = T_r - [W_{node}]$$

11. node→route(E_D)
End

Fig. 3. Algorithm for the Congestion Identification

Fig.4 highlights the schema that has been used in this algorithm.

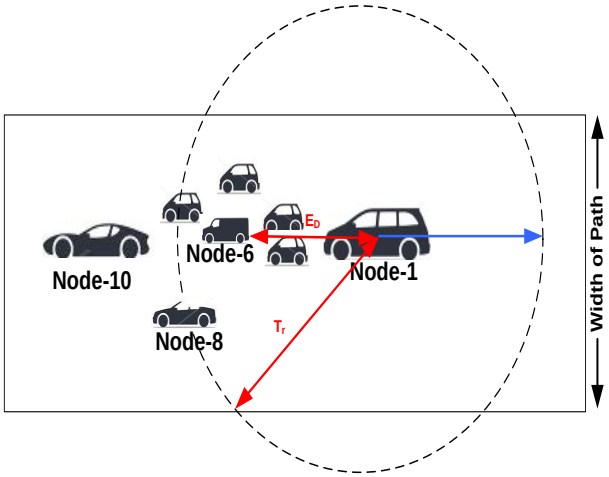


Fig. 4. Schema of routing used in the proposed system

Finally, the algorithm computes effective distance ED (in meters) which is calculated by subtracting transmission range (Tr) with a wavelength of the node (W node)(in meters) (Line-10). The wavelength can be further calculated by highest velocity permissible (in meter per second) on that specific path divided by the frequency of transmission of control message (in Hertz). Fig.4 showcase a scenario which represents node-8 is at more distance from node-1 and hence node-8 will have lesser probability to relay the control message for congestion identification. However, node-6 can be considered to be suitable relay node and will calculate effective distance E_D.

c) Congestion Mitigation:- After the degree of congestion along with location is identified, the next task is to perform mitigation of the congestion. The term mitigation will mean a mechanism to continue communication through alternative routes. From a review of the literature, it has been already seen that existing middleware design of congestion control doesn't consider decision making to be involved. Therefore, we address this problem by formulating an algorithm that supports collaborative networks in hybrid mobile ad-hoc network. The primary target of this algorithm is to minimize the duration of mobile nodes on the path and thereby to recommend the changes in the routes for an alternative solution to congestion.

Algorithm for congestion-free routing

Input: DoC (Degree of Congestion), Th (threshold of congestion), N_N (Neighborhood node), msg (message to estimate congestion), r (response), r_{node} (node with response *r*).

Output: packet forwarding via congestion free routes.

Start

1. If node(DoC)>Th
2. For (N_N=1:max)
3. msg→forward it as alert(N_N)
4. a=number(r)
5. till r_{node}= [node(DoC)>Th]
6. stop
7. select all r_{node} for routing
8. End
9. Repeat step-1-8 until r_{node}= [node(DoC)<Th] in step-5.

End

Fig. 5. Algorithm for congestion-free routing

The algorithm checks for a degree of congestion of the nodes to be more than the threshold level of congestion (Line-1), which corresponds to the state of congestion. In such case, the algorithm looks for all possible neighbor nodes and the search for a chain of neighbor nodes continues (Line-2). In this search, only nodes with a lesser value of the degree of congestion are recorded to be a reliable path for undertaking routing decision. However, for better search optimization, we terminate our search for efficient node exactly for the node with more value of the degree of congestion (Line-5). For all the neighboring nodes whose degree of congestion is within the lower limit of Threshold (Line-9), the control message is transformed to alert message emphasizing on relaying the information that source node is already detected with congestion and hence an alternative congestion-free route must be found.

TABLE II. NOTATION USED IN ALGORITHM DESIGN

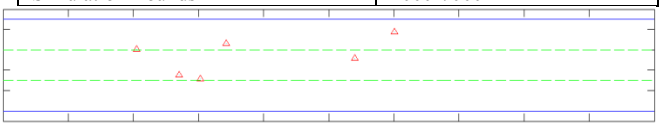
Notation	Meaning
v	velocity
p	Position
t	time interval
d	density of nodes
RBCM	Rule Based Control Matrix
traff	traffic
DoC	Degree of Congestion
F_I	Frequency Intervals
N_N	Number of neighbor nodes
E_D	Effective Distance
T_r	Transmission range
W_{node}	wavelength of node
F	frequency
Th	Threshold
msg	message to estimate congestion
R	response
r_{node}	node with response *r*

VI. RESULT DISCUSSION

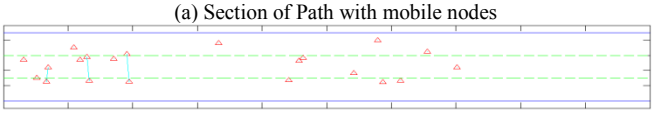
The implementation of the proposed algorithm is carried out in Matlab, where the target was to investigate the factors that have a significant impact on the congestion. A completely new simulator is developed for this purpose considering random mobility factor. We have also taken a case study of a vehicular ad-hoc network to investigate the validity of the proposed concept of middleware system of MIMC. The complete analysis of the results is an outcome of node-to-node communication system considered over the simulation parameters tabulated below:

TABLE III. SIMULATION PARAMETERS

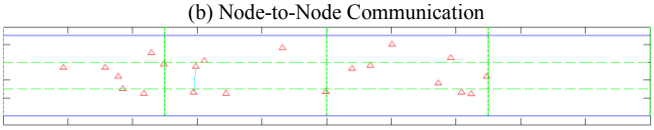
Number of Mobile Nodes	500-1000
Simulation Time	600 Seconds
Minimum velocity	5 meter per second
Maximum velocity	100 meter per seconds
Communication range	200 meters
Simulation Rounds	1000-7000



(a) Section of Path with mobile nodes



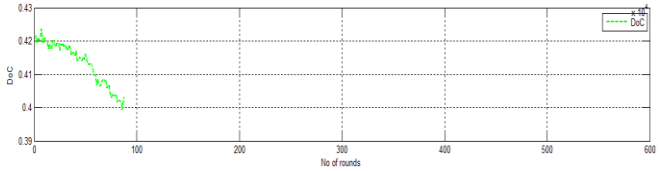
(b) Node-to-Node Communication



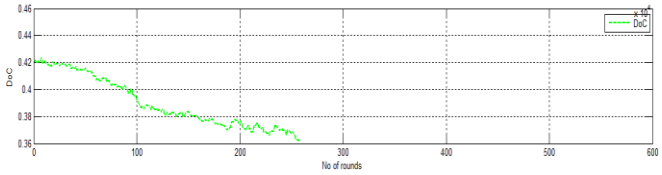
(c) Mobility and Communication in present of obstruction

Fig. 6. MIMC Simulations of Mobile Nodes

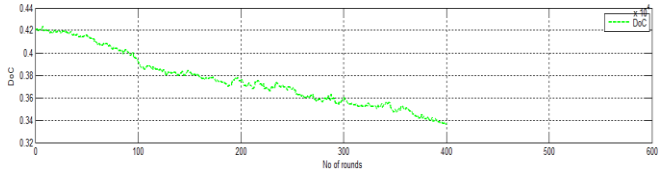
Fig.6 shows the visual outcomes of the simulation carried out in MIMC for the mobile node (Red triangle). A simulation allows a simple segment of the route with multiple lane system (dotted green) as seen in Fig. 6(a). Depending upon the input of simulation parameters (especially the node velocity and transmission range), the node-to-node communication starts between the nodes residing in multiple lanes. The communication link is highlighted as blue line between three pairs of nodes in Fig.6. (b). We also add obstruction in the traffic (shown in vertical dotted green line), which mimics the speed bumpers on the real-word road (Fig.6(c)). It may also represents the traffic signal which we get to see in urban vehicular adhoc networks to some extent. Inclusion of obstruction is meant for reducing the initialized velocity of the mobile node. Although, we initialize minimum and maximum velocity of the node, such initialization is than randomized to total number of mobile nodes considered in the simulation area. This is done in order to incorporate dynamic topology of mobile adhoc network; however we have restricted the dynamicity by letting the mobile nodes to go in one direction of the path. Hence, directionality of the mobile node is fixed in order to check the consistency of the performance of the proposed middleware system with respect to congestion control in real-world road.



(a) Trend of DoC in MIMC during 100th test simulation rounds



(b) Trend of DoC in MIMC during 250th test simulation rounds



(c) Trend of DoC in MIMC during 400th test simulation rounds

Fig. 7. Progressive DoC Observation in MIMC Simulation

The frequently used existing simulators e.g. NS2, OMNeT++, OPNET, etc. does show the simulation of the considered environment, but cannot show the live and dynamic generation of graphical trends during the simulation progress. Dynamic investigation of graphical trend becomes an important operation especially when the investigational rounds are quite high and when there is a need to observe the live trends. We did this feature in our MIMC because we want the implication of MIMC to be more on online result analysis and not on offline result analysis. Hence, this was only possible by Matlab by designing its method and hence highly customizable. Fig.7. (a)-(c) shows the progress being made during each round of simulation. A closer look into the graphical trend will tell that value of degree of congestion i.e. DoC is found to progress in decreasing order.

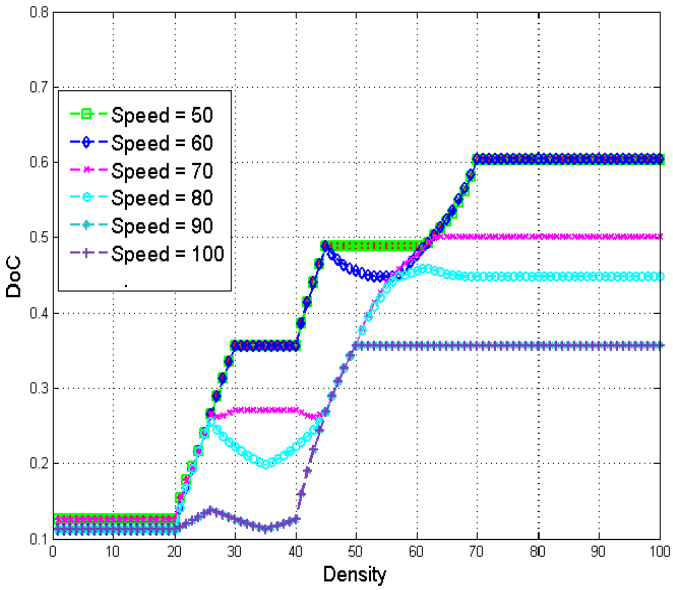
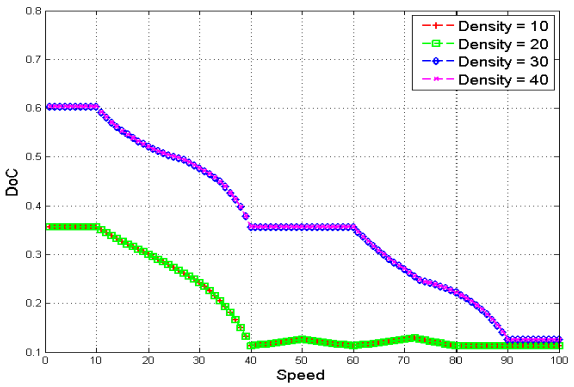
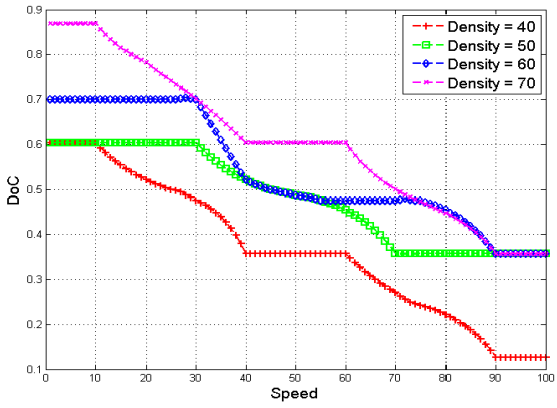


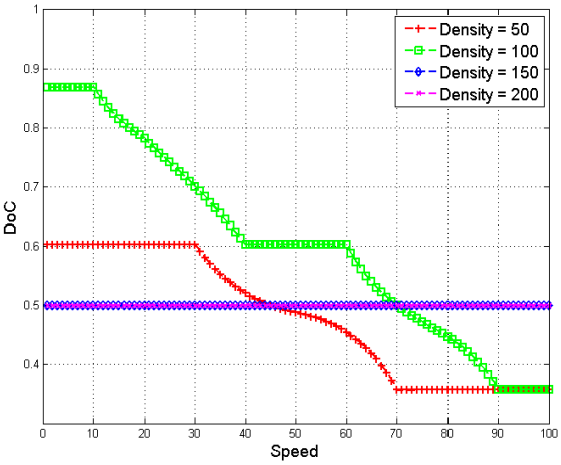
Fig. 8. Impact of Velocity over DoC

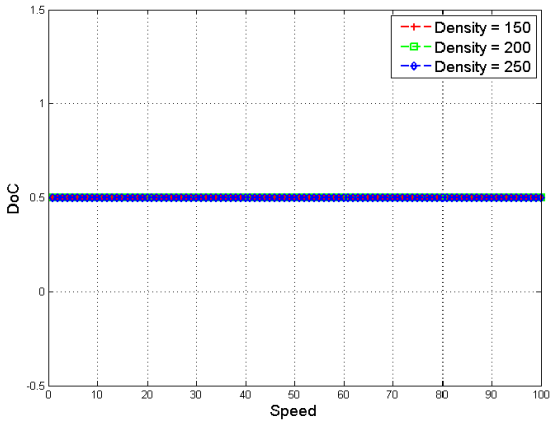
Fig.8 shows the impact of the velocity over the DoC on increasing values of density of nodes. It is already known that increase in density of nodes will eventually increase in DoC value, but there are multiple mobility scenarios that require closer observation. Initially, we initiated the velocity with v=50 ms-1 and kept it constantly increasing with 10ms-1 till it reaches 100ms-1. The outcome shows that MIMC is capable of retaining DoC to lower values even if the speed of the velocity is increased, which is in agreement that increased velocity and reduced density is ideal state of non-congestion and vice-versa.





(a) Density →[10-40 nodes] (b) Density →[40-50 nodes]





(c) Density →[50-200 nodes] (b) Density →[150-250 nodes]

Fig. 9. Impact of Density over DoC

Fig.9 showcases the impact of the node density over the value of the degree of congestion. Anticipating lower value of DoC is quite imperative for the lower value of node density. However, we will like to check the upper limit of it. Hence, the initial investigation with density maintained between 10-40 nodes show a normal decreasing trend of DoC as shown in Fig. 9(a). However, when the density is increased from 50-70 nodes, we find a bit of variation in the DoC trend. It shows increasing values of node density will increase DoC too (Fig. 9(b)). Further, density is again increased from 50 to 200 nodes to find that density of 50, 100, and 150 nodes as a distinctive trend. At the same time, we find a linear behavior for density maintained to be 150 and 200 nodes in Fig.9(c). A similar trend of complete linearity in DoC can be observed when the density is further increased from 150-250. The outcome infers that there are two forms of density factor i.e. i) density within a limit and ii) density beyond a limit. Therefore, density within a limit is considered for 10-150 nodes and density beyond a limit is more than 150 nodes. This inference of the outcome also states that congestion may take place when there is jam of specific numbers of the vehicle that a particular segment of the road can accommodate i.e. Fig. 9(a)-Fig. 9(b). For example, if a segment of a lane can accommodate 70 vehicles so the maximum permissible limit of density can be only 70 here. In such case, the lower limit could be 20-40 nodes depending upon their speed limits.

However, considering the density of more than 150 is nearly impractical for a segment of a road in the real word, and hence the linearity behavior is shown in Fig.9(c) and Fig.9 (d). We carry out this analysis to testify the reliability of outcomes for proposed MIMC.

TABLE IV. NUMERICAL OUTCOME OF COMPARATIVE ANALYSIS

	Delay	Packet Delivery Ratio	Throughput
Vadivel et al. [10]	5.28	1.53	1923
Bhadauria et al.[16]	3.41	1.01	1200
AODV [28]	7.85	2.47	1765
DSDV [29]	7.26	3.12	1632
TORA [30]	6.31	3.23	1771
Proposed	1.89	5.55	4306

The outcome of the proposed system is compared with multiple existing technique with respect to delay, packet delivery ratio, and throughput. The most work carried out by Vadivel et al. [10] is able to perform congestion control but the load balancing algorithm developed by them perform recursive operation of extracting Absolute Congestion Index from their neighbor nodes. The computation increases with increase in node density which was not considered in this study. Similarly work carried out by Bhadauria [16] uses mobile agents instead of middleware. The complete work is done by enhancing AODV itself which cannot contribute much to decision making of routing. Similar forms of problems also exists in conventional AODV, DSDV, and TORA also, which are overcome in proposed MIMC. Similar trend of packet delivery ratio and

throughput can be observed, which shows that proposed system performs better congestion identification in contrast to the existing system. The primary reason behind this outcome is an inclusion of traffic modelling which uses multi-valued logic to accomplish better decision making and better inference to the state of congestion. The secondary reason behind the better outcome of MIMC is we have developed algorithm separately for identification and routing through alternative path which are less congested. This phenomenon results into execution of multiple operations at same time without involving much network related resources. MIMC can be used for any mobile-based adhoc application as well as vehicular application.

VII. Conclusion

This paper has discussed the extension of our prior research work towards middleware design. We strongly believe that adoption of middleware in the hybrid mobile ad-hoc network is extremely important that can also cater up to the need of futuristic communication requirement. After reviewing existing literature, we find that there is few research work of middleware system focusing on addressing congestion problem while existing congestion control protocol doesn't consider middleware system. The proposed MIMC is meant to bridge this gap of research. The significant contribution of MIMC can be summarized as - i) the traffic modelling incorporated in MIMC is capable of representing multiple formats of speed and density to give better inference system. This is carried out to because existing congestion control protocols using frequently used AODV or OLSR lacks decision making for which reason the studies are limited to the offline investigation. ii) The control identification module is designed completely by a particular segment of the route (which represents a small part of any road in real-world). Hence, the applicability of the proposed system in real-world is quite high in the vehicular ad-hoc network. iii) We have not used any sophisticated optimization or mathematical modelling as we aimed to get better response time for any query toward extracting congestion information, iv) the mechanism of routing doesn't include any route acknowledgment, which minimized overhead to a larger extent if multicast protocols would be used here. Finally, the congestion mitigation is done by exploring routes which are less congested. Although the mechanism uses iterative principle, it is highly controlled by the rule-based control matrix. The study outcomes show that MIMC performs better congestion control as compared to the existing system.

References

[1] A. Abdelaziz, M. Nafaa, G. Salim, "Survey of Routing Attacks and Countermeasures in Mobile Ad Hoc Networks", *IEEE International Conference on Computer Modelling and Simulation*, pp.693-698, 2013

[2] W. A. Jabbar, M. Ismail, R. Nordin, S. Arif, "Power-efficient routing schemes for MANETs: a survey and open Issues", *Springer Journal of Wireless Networking*, 2016

[3] D.K Anand, S. Prakash, "A Short Survey of Energy-Efficient Routing Protocols for Mobile Ad-Hoc Networks", *IEEE International Advances in recent Technologies in Communication and Computing*, pp.327-329, 2010

[4] D Maheshwari and R Nedunchezhian, "Load Balancing in Mobile Ad Hoc Networks: A Survey", *International Journal of Computer Applications*, vol.59, iss.16, pp.44-49, December 2012

[5] H. Gupta and P. Pandey, "Survey of routing base congestion control techniques under MANET," *IEEE- International Conference on Emerging Trends in Computing, Communication and Nanotechnology* (ICE-CCN), pp. 241-244, 2013

[6] S. M. Mirhosseini and F. Torgheh, "ADHOCTCP: Improving TCP Performance in Ad Hoc Networks", *IntechOpen*, DOI: 10.5772/13510, 2011

[7] S. M. Adam, R. Hassan, "Delay aware Reactive Routing Protocols for QoS in MANETs: a Review", *Elsevier-ScienceDirect Journal of Applied Research and Technology*, vol.11, Iss.6., 2013

[8] G. Paroux, I. Demeure, D. Baruch, "A survey of middleware for mobile ad hoc networks", *Département Informatique et Réseaux*, 2007

[9] S. Bandyopadhyay, M. Sengupta, S. Maiti, and S. Dutta, "A Survey of Middleware for Internet of Things", *Springer Journal of Recent Trends in Wireless and Mobile Networks*, vol.162, pp.288-296, 2011

[10] R. Vadivel, V. M. Bhaskaran, "Adaptive reliable and congestion control routing protocol for MANET", *Springer Journal of Wireless Network*, 2016

[11] M.D. Sirajuddin, Ch. Rupa and A. Prasad, "Advanced Congestion Control Techniques for MANET", *Springer Journal of Information Systems Design and Intelligent Applications, Advances in Intelligent Systems and Computing*, vol.433, 2016

[12] D. V. Pashchenko, M. S. Jaafar, S. A. Zinkin, D. A. Trokoz, "Directly executable formal models of middleware for MANET and Cloud Networking and Computing", *Journal of Physics: Conference Series*, vol.710, 2016

[13] W. Chen, Q. Guan, S. Jiang, Q. Guan, and T. Huang, "Joint QoS provisioning and congestion control for multi-hop wireless networks", *Springer- EURASIP Journal on Wireless Communications and Networking*, 2016

[14] B.C. Sreenivas, G.C. Bhanu Prakash, and K.V. Ramakrishnan, "M-ADTCP: An Approach for Congestion Control in MANET", *Springer Journal of Advances in Computing & Inf. Technology*, vol.178, pp.531-540, 2013

[15] C. Greco, M. Cagnazzo, B. P. Popescu, "Low-Latency Video Streaming With Congestion Control in Mobile Ad-Hoc Networks", *IEEE Transactions on multimedia*, vol. 14, no. 4, August 2012

[16] S. S. Bhadauria and V. K. Sharma, "Framework and Implimentation of an Agent Based Congestion Control Technique for Mobile Ad-hoc Network", *Springer Journal*, vol.125, pp.318-327, 2011

[17] Z. Qin, L. Iannarioy, C. Giannelliz, P. Bellavista, "MINA: A Reflective Middleware for Managing Dynamic Multinetwork Environments", *IEEE Network Operation and Management Symposium*, 2014

[18] P. Kamisinski, V. Goebel, and T. Plagemann, "A reconfigurable distributed CEP middleware for diverse mobility scenarios", *IEEE International Conference on Pervasive Computing and Communications Workshops*, pp.615-620, 2013

[19] G. Denker, N. Dutt, S. Mehrotra, "Resilient dependable cyber-physical systems: a middleware Perspective", *Springer Journal of Internet Service Application*, vol.3, pp.41-49. 2012

[20] S.G. Pease, "ROAM: supporting safety critical applications in MANETs with cross-layer middleware", *IEEE 14th International Symposium on aWorld of Wireless, Mobile and Multimedia Networks* (WoWMoM), Madrid, Spain, pp.1-2, 2013

[21] P. G. Lopez, R. G. Tinedo, J. M. B. Alsina, "Moving routing protocols to the user space in MANET middleware", *Elsevier- Journal of Network and Computer Applications*, vol.33, pp.588-602, 2010

[22] S. Liu, "A Context-Aware Reflective Middleware Framework for Mobile Ad-hoc and Wireless Sensor Networks", Thesis of Lehigh University, 2012

[23] J.Y. Kim, G. S. Tomar, L. Shrivastava,, "Load Balanced Congestion Adaptive Routing for Mobile Ad Hoc Networks", *International Journal of Distributed Sensor Networks*, 2014

[24] D. W. Kum, W. K. Seo, J. I. Choi and Y. Z. Cho, "Mobility adaptive hybrid routing for mobile ad hoc networks," *IEEE International Conference on Computer Science and Automation Engineering* (CSAE), pp. 377-381, 2012

[25] M. U. Farooq and N. Tapus, "CEHR: Core enabled hybrid routing protocol for mobile ad hoc networks," *IEEE International Conference on*

Intelligent Computer Communication and Processing, pp. 349-354, 2014

[26] P. G. S. Hiremath and C. V. G. Rao, "MERAM: Message exchange with resilient and adaptive middleware system in MANET," *IEEE International Conference on Computational Intelligence and Computing Research* (ICCIC), Madurai, pp. 1-6, 2015

[27] P. G. Sunitha Hiremath, C. V. Guru Rao, "MEEM: A Novel Middleware for Energy Efficiency in Mobile Adhoc Network", *Springer Software Engineering Perspectives and Application in Intelligent Systems*, vol.465, 2016

[28] Perkins, C.; Belding-Royer, E.; Das, S. Ad hoc On-Demand Distance Vector (AODV) Routing. IETF. RFC 3561. Retrieved 2010-06-18., 2003

[29] Perkins, Charles E.; Bhagwat, Pravin "Highly Dynamic Destination-Sequenced Distance-Vector Routing (DSDV) for Mobile Computers" (pdf). Retrieved 2006-10-20, 2994

[30] V. D. Park and M. S. Corson. "A highly adaptive distributed routing algorithm for mobile wireless networks". In Proceedings of INFOCOM, 1997

Statistical Implicative Similarity Measures for User-based Collaborative Filtering Recommender System

Nghia Quoc Phan
Testing Office,
Travinh University,
Travinh City, Vietnam

Phuong Hoai Dang
Information Technology Faculty,
Danang University of Science and Technology, Danang City, Vietnam

Hiep Xuan Huynh
College of Information & Communications Technology, Cantho University,
Cantho City, Vietnam

***Abstract*—This paper proposes a new similarity measures for User-based collaborative filtering recommender system. The similarity measures for two users are based on the Implication intensity measures. It is called statistical implicative similarity measures (SIS). This similarity measures is applied to build the experimental framework for User-based collaborative filtering recommender model. The experiments on MovieLense dataset show that the model using our similarity measures has fairly accurate results compared with User-based collaborative filtering model using traditional similarity measures as Pearson correlation, Cosine similarity, and Jaccard.**

Keywords—Similarity measures; Implication intensity; User-based collaborative filtering recommender system; statistical implicative similarity measures

I. INTRODUCTION

Currently, recommender system is considered as a useful tool for solving partial information overload of the Internet [3][12]. Its development is always associated with the development of web technologies and machine learning algorithms. Based on the method of collecting and processing the data, the recommender systems can divide into three generations. The first generation of recommender systems uses traditional websites to gather information from three sources: (1) content-based data from the purchase or the use of products and services; (2) demographic data selected from the customer profile; (3) memory-based data collected from the user's preferences. In this generation, the quality of recommendation results is improved based on data classification algorithms and the integration of the data classification algorithms [3][15][27]. The second generation of recommender systems is the increasing use of Web 2.0 by collecting information through social network like Facebook, Zalo and other social networking sites. To satisfy explosive information issue from social networking sites, this generation continues to develop and improve the existing integrated methods and enhance solutions to exploit information from social networks more efficiently such as trust-aware algorithms [20], social adaptive approaches [4], social networks analysis [12][28] and other methods. The third generation of recommender systems is developed in parallel with the web 3.0 with information collected from integrated devices on the Internet such as cameras, sensors [22]. This generation uses approaches to integrate location information into the available recommendation algorithms in order to broaden its application in various fields such as health, weather, environment, and universe [1].

User-based collaborative filtering recommender system is the first version of the recommender systems based on collaborative filtering. It was first introduced in the article "GroupLens: an open architecture for Collaborative filtering of Netnews" in 1994 for GroupLens Usenet recommender system [21]. Subsequently, there are two other recommender systems also use this recommendation method: one for users to listen to music Ringo [25] and the other for users to watch movies Bellcore [26]. User-based collaborative filtering recommender system is a simple algorithm to clarify the core premise of collaborative filtering methods. That is to find out users in the past who had the same behavior as current users. Then, the value rating of users for the items is used to predict the preferences of current users. Thus, in order to obtain a list of items to introduce to new users, User-based collaborative filtering recommender system requires a function to compute the similarity of two users and a method to calculate the average deviation of rating values of similar users based on a rating matrix of users for items [14][15][17].

From the first appearance with the name "The information Lense system" in 1987 [13], recommender system has been developed greatly in technology and its application in the fields of life. In particular, recommender systems are used by many managers as an effective tool in order to support business activities in various fields such as Amazon, Netflix, and Pandora [2]. However, the present generation of recommender systems has not fully met the requirements of users yet. Therefore, research on recommender systems continues to be concerned such as research to improve methods and algorithms to increase accuracy of the existing recommender model [11][18][24], research to improve recommender systems to adapt to the information explosion and research to propose a new recommender model [8]. In addition, some new research directions are also set out, such as research on proper combination of existing recommendation methods that use different types of available information; research on using the maximum capabilities of the sensors and devices on the Internet; research on collecting and integrating information on trends related to habits, consumption and individual tastes of users in the recommendation process; research on ensuring the security conditions and privacy in the entire process of recommendation system; research on proposing the measures for evaluating recommender systems and develop a standard for assessment measures and research on developing a framework for automated analysis on heterogeneous data.

In this paper, a new similarity measures between two users based on Implication intensity measures is proposed for User-based collaborative filtering recommender system. We describe how to build measures and their application in User-based collaborative filtering recommender model. After building the model, the experiments of model was conducted on MovieLense dataset [5] and compared the results with the User-based collaborative filtering recommender model that uses the traditional similarity measures such as Pearson, Cosine, and Jaccard.

This paper has six sections. Section 1 introduces general recommender systems, User-based collaborative filtering recommender system, relevant studies, and addressing the research issue. Section 2 shows how to build a similarity measures between two users based on the statistical implication intensity measures. Section 3 describes the required steps to build User-based collaborative filtering recommender model based on statistical implicative similarity measures. Section 4 presents the evaluation methods of recommender systems. Section 5 presents the experimental results of the model and compares the results with other models. The final section summarizes some importantly achieved results of model using similarity measures between two users based on Implication intensity measures

II. Similarity Measures Between Two Users Based on Implication Intensity Measures

A. Implication intensity measures

Statistical implicative analysis is the method of data analysis studying implicative relationships between variables or data attributes, allowing detecting the asymmetrical rules A → B in the form "if A then that almost B" or "consider to what extent that B will meet implication of A" [23].

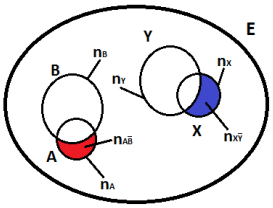


Fig. 1. The model represents a statistical implication rule A → B

Every rule A → B is represented by a table based on the concept of probability called the probability distribution table 2x2 in order to store the counting frequency in the satisfaction of the established conditions. From this probability distribution table, the probability value is calculated based on the frequency of n, n_A, n_B, n_{AB}, $n_{A\bar{B}}$ respectively as follows: $P(A) = \frac{n_A}{n}$, $P(B) = \frac{n_B}{n}$, $P(A \cap B) = \frac{n_{AB}}{n}$, $P(A \cap \bar{B}) = \frac{n_{A\bar{B}}}{n}$.

TABLE I. Probability Distribution Table 2x2 of Statistical Implication Rule A → B

	A → B		
	B	$\bar{B}$	
A	n_{AB}	$n_{A\bar{B}}$	n_A
$\bar{A}$	$n_{\bar{A}B}$	$n_{\bar{A}\bar{B}}$	$n_{\bar{A}}$
	n_B	$n_{\bar{B}}$	n

Interestingness value of Implication intensity measures for rule A → B is determined by a formula based on the probability distribution table 2x2 under the following form [23]:

$$\Pr(card(X \cap \bar{Y}) \leq card(A \cap \bar{B}) = \sum_{s=0}^{card(A \cap \bar{B})} \frac{\lambda^s}{s!} e^{-\lambda}$$

Where $\lambda = \frac{n_A(n-n_B)}{n}$

B. Building the similarity measures between two users

In order to calculate the similarity between two users based on Implication intensity measures, the algorithm is proposed consisting of the following steps:

Input: Rating data for items of two users u, v.

Output: Similarity values between two users u, v.

Begin

Step 1: Select the statistical implication rules for two users

- Select the Items that user u rated (I_u);
- Select the Items that user v did not rate (I_v);
- For rule set to be generated from rating matrix of users:
  Begin
  Select the rules of the form $\{X\} \rightarrow \{Y\}$
  where $X \in I_u$; $Y \in I_v$ and $X \cap Y = \emptyset$;
  End;

Step 2: Count the parameters $n, n_A, n_B, n_{A\bar{B}}$

- For each rule in the selected rule set:
  Begin
  Count the parameters n, n_A , n_B, $n_{A\bar{B}}$;
  End;

Step 3: Calculate Implication intensity value for rule set

- For each rule in the selected rule set:
  Begin
  Calculate the value of Implication intensity measures: Implicationintensity$(n, n_A, n_B, n_{A\bar{B}})$;
  End;

Step 4: Calculate the similarity between two users

$$SIS(u,v) = 1 - \left(\frac{\sum_{i=1}^{k} \text{Implicationintensity}_{I_{u,i} \rightarrow I_{v,i}}(n, n_A, n_B, n_{A\bar{B}})}{k} \right)$$

End;

Example: Let us a rating matrix of two users who rated for 4 items as follows:

	i_1	i_2	i_3	i_4
u_1	0	4	4	1
u_2	2	0	4	0

At the first step, selecting statistical implication rules between user u_1 and user u_2 including:

N^0	Statistical implication rules
1	{V2=4} => {V1=0}
2	{V4=1} => {V1=0}
3	{V3=4} => {V1=0}
4	{V3=4} => {V2=0}
5	{V3=4} => {V4=0}
6	{V2=4,V4=1} => {V1=0}
7	{V2=4,V3=4} => {V1=0}
8	{V3=4,V4=1} => {V1=0}
9	{V2=4,V3=4,V4=1} => {V1=0}

At the next step, count the parameters $n, n_A, n_B, n_{A\bar{B}}$ for each statistical implication rule and calculate implication intensity values based on the parameters:

N^0	Statistical implication rules	n	n_A	n_B	$n_{A\bar{B}}$	ImplicationIntensity
1	{V2=4} => {V1=0}	2	1	1	0	0.486582881
2	{V4=1} => {V1=0}	2	1	1	0	0.486582881
3	{V3=4} => {V1=0}	2	2	1	1	0.384940011
4	{V3=4} => {V2=0}	2	2	1	1	0.384940011
5	{V3=4} => {V4=0}	2	2	1	1	0.384940011
6	{V2=4,V4=1} => {V1=0}	2	1	1	0	0.486582881
7	{V2=4,V3=4} => {V1=0}	2	1	1	0	0.486582881
8	{V3=4,V4=1} => {V1=0}	2	1	1	0	0.486582881
9	{V2=4,V3=4,V4=1} => {V1=0}	2	1	1	0	0.486582881

At the final step, the similarity between user u_1 and user u_2 is determined as follows:

$$SIS(u_1, u_2) = 1 - 0.452701924 = 0.54729808.$$

III. User-based Collaborative Filtering Recommender System based on Sis Measures

User-based collaborative filtering recommender model based on statistical implicative similarity measures is defined as follows:

Suppose that $U = \{u_1, u_2, \dots, u_m\}$ is a set of m users, $I = \{i_1, i_2, \dots, i_n\}$ is a set of n items, $R = \{r_{j,k}\}$ is a rating matrix of m users for n items with each row representing a user u_j $(1 \geq j \geq m)$, each column represents an item i_k $(1 \geq k \geq n)$, $r_{j,k}$ is the rating value of user u_j for item i_k and $u_a \in U$ is user who needs recommendation.

According to initial data, the model implemented through the following steps:

Step 1: Measure the similarity between users u_a and other users in the system by using function: $SIS(u_a, u_i)$.

Step 2: Determine the list of k similarity users who are similar with u_a: $N(a) \in U$.

Step 3: Identify the item categories that user u_a has not rated yet, determined by: $I_a = I \backslash \{i_k \in I | r_{a,k} \geq 1\}$ and calculate predicted rating values for this item categories by the following formula: $\hat{r}_{a,k} = \frac{1}{\sum_{i \in N(a)} s_{a,i}} \sum_{i \in N(a)} s_{a,i} r_{i,k}$, where $s_{a,i}$ is the similarity value between user u_a and user u_i.

Step 4: Recommend N-items which obtained the highest predicted rating value to user u_a: $T_N \in I_a$.

The model is presented by the following diagram:

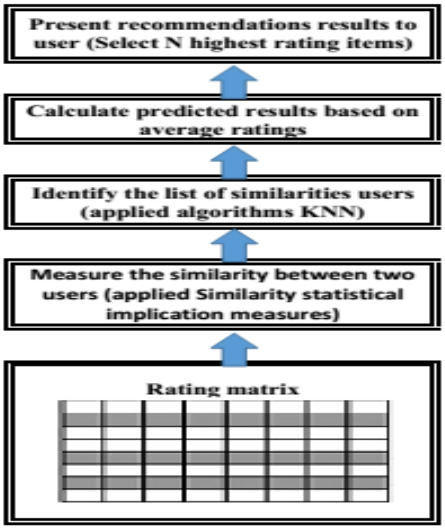


Fig. 2. User-based collaborative filtering recommender model based on statistical implicative similarity measures

IV. Evaluating the Recommender Model

The evaluation of the accuracy of the recommender model is an important step in the recommender system design process [6][7][9]. It helps designers choose models, check the accuracy of the model before applying the model into practice. To evaluate User-based collaborative filtering recommender model, the recommender system designers can be conducted through two steps:

A. Preparing the data to evaluate the models

In order to evaluate the quality of a predictive model, experimental datasets have divided into two parts: one for modeling and the rest for testing. Therefore, the first step is to prepare the data; in this step the experimental dataset is divided into two subsets: training dataset and testing dataset [17]. Currently, many methods are being used to split datasets for evaluating recommender models such as:

Splitting: is the initial method to build a training set and test set by cutting experimental dataset into 2 parts [17]. For this method, the model designer should decide the percentage for the training set and test set. For example, the training set accounts for 80 percent and the test set account for the remaining 20 percent.

Bootstrap sampling: is a method used to build a training set and test set by cutting the experimental dataset into 2 parts. However, this approach is done randomly and repeatedly in order that a user may be a member of the training set in this cutting time but is a member of test set in the next cutting time. This can overcome the disadvantages of heterogeneity of the experimental dataset and increase optimization for small-sized dataset. [17].

K-fold cross-validation: is a method used to build a training set and test set by cutting the experimental dataset into

k subsets with the same size (called k-fold). After that, the model is evaluated k times. Every evaluation uses one subset for the test set and the k-1 subsets are used as the training set. The evaluation results of this method are average value of k evaluations. This approach ensures that all users have appeared at least one time in the test set [17]. Therefore, it is the most accurate of the three methods. However, it is costly for the calculation compared with the remaining two methods.

B. Evaluate recommender model

There are two methods for evaluating recommender model: evaluation based on the ratings and evaluation based on the recommendations. The first method evaluates the ratings generated by the model. The remaining method evaluates directly on the recommendations of the model.

Evaluation based on the ratings: a method evaluates the accuracy of the model by comparing the predicted rating value with the real value. More precisely, this method is to find out the average error value based on three indicators RMSE, MSE and MAE. A model is evaluated good if these indicators show low value [7][9].

Root mean square error (RMSE): This is the standard deviation between the real and predicted ratings.

$$RMSE = \sqrt{\frac{\sum_{(i,j)\in\kappa}(r_{ij} - \hat{r}_{ij})^2}{|\kappa|}}$$

Mean squared error (MSE): This is the mean of the squared difference between the real and predicted ratings. It's the square of RMSE, so it contains the same information.

$$MSE = \frac{\sum_{(i,j)\in\kappa}(r_{ij} - \hat{r}_{ij})^2}{|\kappa|}$$

Mean absolute error (MAE): This is the mean of the absolute difference between the real and predicted ratings.

$$MAE = \frac{1}{|\kappa|}\sum_{(i,j)\in\kappa}\left|r_{ij} - \hat{r}_{ij}\right|$$

where κ is the set of all user ratings for items; r_{ij} real rating value of user i for item j; $\hat{r}_{ij}$ is predicted rating value of user i for item j.

Evaluation based on the recommendations: a method evaluates the accuracy of the model by comparing the model's recommendations to purchase choice of the users. This approach uses confusion matrix to calculate the value of three indicators: Precision, Recall and F-measure. The model is evaluated good if three indices gain high value [7][9].

TABLE II. CONFUSION MATRIX

User Choices	Recommendations of the model	
	Recommend	Not recommend
Purchase	TP	FN
Not purchase	FP	TN

Let's explains the confusion matrix:

True Positives (TP): These are recommended items that have been purchased.

False Positives (FP): These are recommended items that haven't been purchased.

False Negatives (FN): These are not recommended items that have been purchased.

True Negatives (TN): These are not recommended items that haven't been purchased.

The formula of three indicators is used to evaluate:

$$\text{Precision} = \frac{Correctly\ recommended\ items}{Total\ recommended\ items} = \frac{TP}{TP + FP}$$

$$\text{Recall} = \frac{Correctly\ recommended\ items}{Total\ useful\ recommendations} = \frac{TP}{TP + FN}$$

$$F - measure = \frac{2 * \text{Precision} * \text{Recall}}{\text{Precision} + \text{Recall}}$$

V. EXPERIMENT

A. Data description

The experimental dataset is MovieLense [5] of GroupLens research project at the University of Minnesota in 1997. This dataset is collected from the rating results of 943 users for 1.664 movies (99.392 rating results from 0 to 5) through the MovieLense website (movielens.umn.edu) during 7 months (from 09/19/1997 to 22/04/1998). This dataset is organized in a matrix format consisting of 943 rows, 1.664 columns and 1.569.152 cells containing rated values. However, each user is able to watches her/his favourite movies. Thus, the rating matrix has only 99.392 rating values of users for movie categories.

B. Implementation tools

In order to conduct experiment, we use ARQAT tool which is developed on language R by our team. This is a tool package to be developed from engine platform ARQAT on language Java [10]. This tool includes the following functions: processing data, generating statistical implication rules, counting parameters $n, n_A, n_B, n_{A\bar{B}}$, calculating value of objective interestingness measures based on 4 statistical implication parameters, calculating similarity of two users based on statistical implicative similarity measures, and designing and evaluation recommender models [16].

C. Select and process data

The MovieLense dataset is stored under a real rating matrix. It consists of 943 rows, 1.664 columns and 1.569.152 cells containing rated value. In particular, more than 93 percent cells have rating values equal 0 and nearly 7 percent remaining cells have rating values from 1 to 5 (value 0 is 1.469.760; value 1 is 6.059 ; value 2 is 11.307; value 3 is 27.002; value 4 is 33.947; value 5 is 21.077). Therefore, the entire MovieLense dataset has only truly 99.392 rating value from users for movies. In particular, the majority of rating values range from 3 to 5 and 4 is rating value with the highest amount. In order to find out the number of users rating for each movie and the number of movies that each user rated, statistical calculations are performed on each movie and each user and illustrate the results in Figure 3.

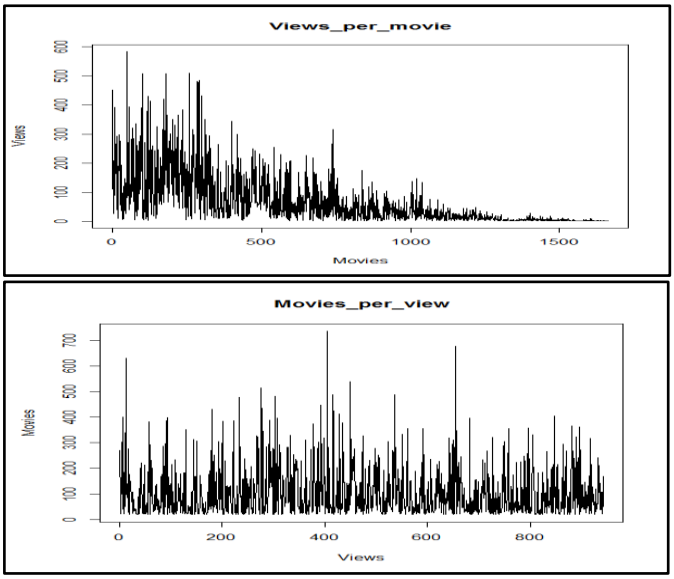


Fig. 3. The chart presents statistical results for each movie and each user on MovieLense dataset

Figure 3 reflects that some movies have only been rated by a few users and some users have only been rated for a few movies. If this case is used for training model, it is likely to lead to bias due to lack of data. Thus, users rate at least for 50 movies and movies rated by at least 100 users are selected to build experimental datasets for model. From there, rating matrix has only 560 rows, 332 columns and 55.298 rating value. In particular, the dataset is split into two subsets: Training set accounting for 80 percent and Test set does the remaining 20 percent.

D. The result of the model

From the result of data processing steps, the model trains on training set with 445 users and tests on test set with 115 users. The result of the model is exported in matrix format with structure 6 x 115 (each column is a user; each cell is a selected movie to recommend for the user in the corresponding column). Figure 4 presents the results of recommender model to the first 4 users; each of them selects the 6 highest rated movies.

```
     [,1]                                [,2]
[1,] "Star Wars (1977)"                  "Citizen Kane (1941)"
[2,] "Godfather, The (1972)"             "2001: A Space Odyssey (1968)"
[3,] "Shawshank Redemption, The (1994)"  "Schindler's List (1993)"
[4,] "Raiders of the Lost Ark (1981)"    "Vertigo (1958)"
[5,] "Good Will Hunting (1997)"          "Casablanca (1942)"
[6,] "Clockwork Orange, A (1971)"        "Blade Runner (1982)"
     [,3]                                [,4]
[1,] "Shawshank Redemption, The (1994)"  "Secrets & Lies (1996)"
[2,] "Schindler's List (1993)"           "Good Will Hunting (1997)"
[3,] "GoodFellas (1990)"                 "Sense and Sensibility (1995)"
[4,] "Usual Suspects, The (1995)"        "Welcome to the Dollhouse (1995)"
[5,] "People vs. Larry Flynt, The (1996)" "Silence of the Lambs, The (1991)"
[6,] "L.A. Confidential (1997)"          "Lone Star (1996)"
```

Fig. 4. Presenting recommendation results of the first 4 users

Based on the recommendation result matrix, we calculate the number of times that each movie is recommended and build a histogram for the distribution of movies in Figure 5. The chart shows that the number of movies is recommended from 5 times or less accounting for relatively large numbers. In particular, up to 38 movies are only recommended 1 time and 24 movies are recommended twice. In contrast, the number of movies is recommended from 5 to 40 times accounting for a very small number. Most of them have the number from 1 to 2 movies.

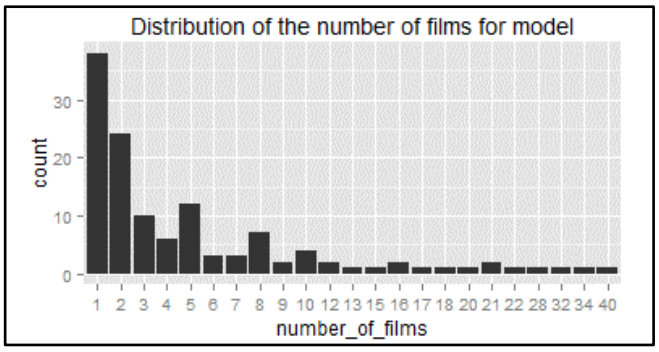


Fig. 5. Presentation distribution of the number of movies for model

E. Evaluation the model

1) Evaluation based on the ratings

In this section, the error parameters (RMSE, MSE, MAE) are calculated for each user and for the model based on the data which is built by k-fold method (with k = 4). For the error parameters of each user, the distribution of each error parameter is performed by a chart and compared them with the error parameters of the model using similarity Pearson measures (Figure 6). The chart shows that the number of users distributed on the error parameters of the model using SIS measures has a higher value than that of the model using similarity Pearson measures. For the error parameters of the model, the value of error parameters is compared with the error parameters of the model using Pearson similarity measures in table 3. The results of comparison found that the values of error parameters of our model are lower than the model using similarity Pearson measures on MovieLense dataset.

TABLE III. PRESENT COMPARISON ERROR PARAMETERS OF TWO MODELS

	RMSE	MSE	MAE
Model using SIS measures	0.9146675	0.8366166	0.7132866
Model using similarity Pearson measures	0.9796664	0.9597462	0.7704055

Model using SIS measures **Model using similarity Pearson measures**

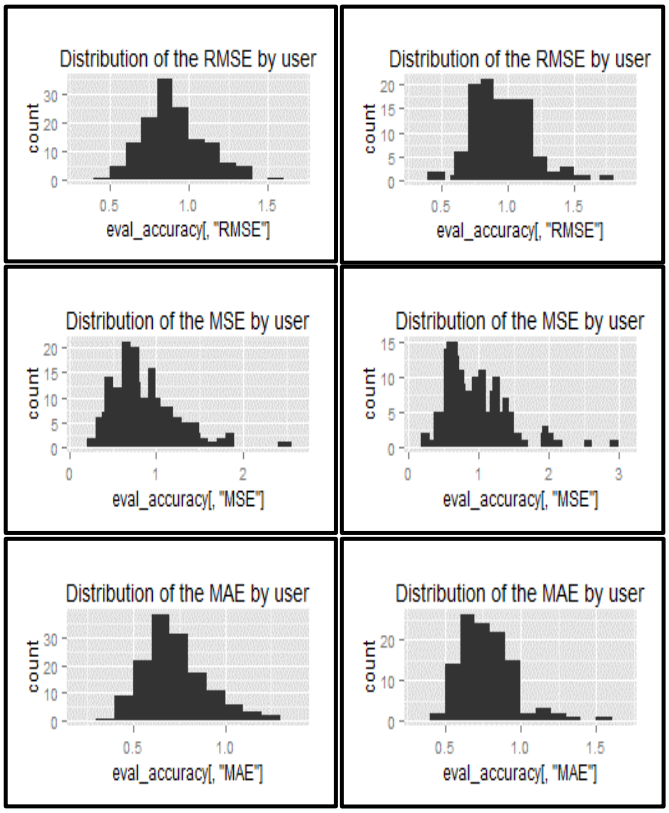


Fig. 6. Comparison of error parameters of each user on two models

2) Evaluation based on the recommendations

In this evaluation method, we also calculate the indicators: TP, FP, FN, TN, Precision, Recall and F-measure based on k-fold data that has been built above. In order to examine the accuracy of the model, the model is tested with the number of movies to be recommended to users which gradually increase (1 to 15). The average rating result of 4 k-fold on models that uses statistical implicative similarity measures and the model using similarity Pearson measures is presented in Figure 7. This figure shows that the indicators Precision, Recall and F-measure on both models are of relatively equal value. This shows that the model that uses statistical implicative similarity measures has the same accuracy as the model that uses similarity Pearson measures on MovieLense dataset.

Model using SIS measures

	TP	FP	FN	TN	precision	recall	f-measure
1	0.7053571	0.2946429	72.43571	243.5643	0.7053571	0.01226171	0.02410440
2	1.3250000	0.6750000	71.81607	243.1839	0.6625000	0.02248717	0.04349790
3	1.9160714	1.0839286	71.22500	242.7750	0.6386905	0.03234240	0.06156713
4	2.4607143	1.5392857	70.68036	242.3196	0.6151786	0.04137372	0.07753297
5	3.0125000	1.9875000	70.12857	241.8714	0.6025000	0.05031998	0.09288254
6	3.5250000	2.4750000	69.61607	241.3839	0.5875000	0.05844200	0.10630885
7	4.0285714	2.9714286	69.11250	240.8875	0.5755102	0.06632461	0.11894179
8	4.4821429	3.5178571	68.65893	240.3411	0.5602679	0.07333510	0.12969415
9	4.9642857	4.0357143	68.17679	239.8232	0.5515873	0.08061004	0.14066328
10	5.4017857	4.5982143	67.73929	239.2607	0.5401786	0.08729768	0.15030477
11	5.8607143	5.1392857	67.28036	238.7196	0.5327922	0.09398769	0.15978786
12	6.3125000	5.6875000	66.82857	238.1714	0.5260417	0.10077316	0.16914367
13	6.7357143	6.2642857	66.40536	237.5946	0.5181319	0.10666154	0.17690566
14	7.1607143	6.8392857	65.98036	237.0196	0.5114796	0.11400974	0.18645771
15	7.5482143	7.4517857	65.59286	236.4071	0.5032143	0.11933871	0.19292476

Model using similarity Pearson measures

	TP	FP	FN	TN	precision	recall	f-measure
1	0.6696429	0.3303571	72.45179	243.5482	0.6696429	0.01107190	0.02178363
2	1.3017857	0.6982143	71.81964	243.1804	0.6508929	0.02151953	0.04166166
3	1.8875000	1.1125000	71.23393	242.7661	0.6291667	0.03130096	0.05963508
4	2.4482143	1.5517857	70.67321	242.3268	0.6120536	0.04036936	0.07574293
5	3.0035714	1.9964286	70.11786	241.8821	0.6007143	0.04897361	0.09056394
6	3.5392857	2.4607143	69.58214	241.4179	0.5898810	0.05732153	0.10448934
7	4.0232143	2.9767857	69.09821	240.9018	0.5747449	0.06465853	0.11624010
8	4.5232143	3.4767857	68.59821	240.4018	0.5654018	0.07248945	0.12850361
9	4.9821429	4.0178571	68.13929	239.8607	0.5535714	0.07909891	0.13841931
10	5.4571429	4.5428571	67.66429	239.3357	0.5457143	0.08603460	0.14863599
11	5.8803571	5.1196429	67.24107	238.7589	0.5345779	0.09229024	0.15740574
12	6.3053571	5.6946429	66.81607	238.1839	0.5254464	0.09832497	0.16565204
13	6.7446429	6.2553571	66.37679	237.6232	0.5188187	0.10567796	0.17559006
14	7.1589286	6.8410714	65.96250	237.0375	0.5113520	0.11138375	0.18292286
15	7.5767857	7.4232143	65.54464	236.4554	0.5051190	0.11765170	0.19085069

Fig. 7. Comparison of indicators based on the recommendations of two models

VI. Conclusion

In this paper, we built User-based collaborative filtering recommender model by suggesting a new similarity measures based on Implications intensity measures in order to determine the similarity of two users. Like other User-based collaborative filtering recommender models, our model follows the main steps such as process data, build the rating matrix, compute the similarity between two users, identify the item list that the similarity users rated highly in order to the recommendation results and evaluate accuracy of the model. However, the new point of this model is to identify the similarity user list, using statistical implicative similarity measures instead of using the familiar measures such as Pearson correlation, Cosine similarity, Jaccard to determine the similarity between two users. The experiments show that our model results are relatively accurate on MovieLense dataset. In particular, the error parameters (RMSE, MSE, MAE) have a lower value than the model using the similarity Pearson measure; Indicators of Precision, Recall and F-measure have the equivalent values compared to the model using Pearson similarity measures. This result shows that the User-based collaborative filtering recommender model using the statistical implicative similarity measures is capable to practice.

References

[1] Ali Elkahky, Yang Song and Xiaodong He, "A Multi-View Deep Learning Approach for Cross Domain User Modeling in Recommendation Systems," International World Wide Web Conference Committee (IW3C2), WWW 2015, May 18–22, 2015, Florence, Italy, ACM 978-1-4503-3469-3/15/05, 2015.

[2] Ben Schafer, Joseph Konstan and John Ried, "Recommender Systems in E-Commerce," EC '99 Proceedings of the 1st ACM conference on Electronic commerce, ISBN:1-58113-176-3, 1999, pp.158-166.

[3] Bobadilla, Ortega, Hernando and Gutiérrez, "Recommender systems survey," Knowledge-Based Systems 46 (2013), 2013, pp.109–132.

[4] F. Liu and H. J. Lee, "Use of social network information to enhance collaborative filtering performance," Expert Systems with Applications 37(7), 2010, pp.4772-4778.

[5] F. Maxwell Harper and Joseph A. Konstan, "The MovieLens Datasets: History and Context," ACM Transactions on Interactive Intelligent Systems (TiiS) 5, 4, Article 19, 2015, pp.1-19.

[6] Feng Zhang, TiGong, Victor E. Lee, Gansen Zhao, Chunming Rong and Guangzhi Qu, "Fast algorithms to evaluate collaborative filtering recommender systems," Knowledge-Based Systems 96 (2016), 2016, pp.96–103.

[7] Gunawardana A, Shani G, "A Survey of Accuracy Evaluation Metrics of Recommendation Tasks," Journal of Machine Learning Research, 10, 2009, p.2935–2962.

[8] Hao Wang, Naiyan Wang, and Dit-Yan Yeung, "Collaborative Deep Learning for Recommender Systems," KDD'15, August 10-13, 2015, Sydney, NSW, Australia, 2015 ACM, ISBN 978-1-4503-3664-2, DOI: http://dx.doi.org/10.1145/2783258.2783273, 2015, pp.1235-1244.

[9] Herlocker JL, Konstan JA, Terveen LG, and Riedl JT, "Evaluating collaborative filtering recommender systems," ACM Transactions on Information Systems, 22(1), ISSN 1046-8188, 2004, pp.5–53.

[10] Hiep Xuan Huynh, Fabrice Guillet and Henri Briand, "ARQAT: An Exploratory Analysis Tool For Interestingness Measures", in International symposium on Applied Stochastic Models and Data Analysis, 2005, pp.334-344.

[11] Huizhi Liang and Timothy Baldwin, "A Probabilistic Rating Auto-encoder for Personalized Recommender Systems," CIKM'15, October 19–23, 2015, Melbourne, Australia, 2015 ACM, ISBN 978-1-4503-3794-6, DOI: http://dx.doi.org/10.1145/2806416.2806633, 2015, pp.1863-1866.

[12] Jiliang Tang, Suhang Wang, Xia Hu, Dawei Yin, Yingzhou Bi, Yi Chang and Huan Liu, "Recommendation with Social Dimensions," Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence (AAAI-16), 2016, pp.251-257.

[13] Malone TW, Grant KR, Turbak FA, and Brobst SA, "Cohen MD Intelligent information sharing systems," Communications of the ACM, 30(5), ISSN 0001-0782, 1987, pp.390–402.

[14] Martin P. Robillard, Walid Maalej, Robert J. Walker and Thomas Zimmermann, "Recommendation Systems in Software Engineering," Springer Heidelberg New York Dordrecht London, ISBN 978-3-642-45135-5, 2014.

[15] Michael D. Ekstrand, John T. Riedl and Joseph A. Konstan, "Collaborative Filtering Recommender Systems," Foundations and Trends in Human–Computer Interaction Vol. 4, No. 2 (2010), 2010, pp.81–173.

[16] Michael Hahsler, "Lab for Developing and Testing Recommender Algorithms," Copyright (C) Michael Hahsler (PCA and SVD implementation (C) Saurabh Bathnagar), http://R-Forge.R-project.org/projects/recommenderlab/, 2015.

[17] Michael Hahsler, "recommenderlab: A Framework for Developing and Testing Recommendation Algorithms," the Intelligent Data Analysis Lab at SMU, http://lyle.smu.edu/IDA/recommenderlab/, 2011.

[18] Mingjie Qian, Liangjie Hong, Yue Shi and Suju Rajan, "Structured Sparse Regression for Recommender Systems," CIKM'15,October 19–23, 2015, Melbourne, VIC, Australia, 2015 ACM, ISBN 978-1-4503-3794-6, DOI: http://dx.doi.org/10.1145/2806416.2806641, 2015, pp.1895-1898.

[19] Nghia Quoc Phan, Hiep Xuan Huynh, Fabrice Guillet and Régis Gras, "Classifying objective interestingness measures based on the tendency of value variation," VIII Colloque International –VIII International Conference, A.S.I. Analyse Statistique Implicative — Statistical Implicative Analysis Radès (Tunisie) - Novembre 2015, Bibliotheeeefque Nationale de Tunisie, ISBN: 978-9973-9819-0-5, 2015, pp.143-172.

[20] P. Bedi, H. Kaur, and S. Marwaha, "Trust based recommender system for semantic web," IJCAI'07 - Proceedings of the 2007 International Joint Conferences on Artificial Intelligence, 2007, pp.2677-2682.

[21] P. Resnick, N. Iacovou, M. Suchak, P. Bergstrom, and J. Riedl, "GroupLens: an open architecture for collaborative filtering of netnews," inACM CSCW '94, 1994, pp. 175–186.

[22] Quanjun Chen, Xuan Song, Harutoshi Yamada and Ryosuke Shibasaki, "Learning Deep Representation from Big and Heterogeneous Data for Traffic Accident Inference," Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence (AAAI-16), 2016, pp.338-344.

[23] R. Gras and P. Kuntz, "An overview of the Statistical Implicative Analysis (SIA) development," Statistical Implicative Analysis – Studies in Computational Intelligence (Volume 127), Springer-Verlag, 2008, pp.11-40.

[24] Ting Yuan, Jian Cheng, Xi Zhang, Shuang Qiu, and Hanqing Lu, "Recommendation by Mining Multiple User Behaviors with Group Sparsity," AAAI Publications, Twenty-Eighth AAAI Conference on Artificial Intelligence, 2014, pp.222-228.

[25] U. Shardanand and P. Maes, "Social information filtering: Algorithms for automating "word of mouth," ACM Press/Addison-Wesley Publishing Co., 1995, pp. 210–217.

[26] W. Hill, L. Stead, M. Rosenstein, and G. Furnas, "Recommending and evaluating choices in a virtual community of use," ACM Press/Addison-Wesley Publishing Co., 1995, pp. 194–201.

[27] Xiaoyuan Su and Taghi M. Khoshgoftaar, "A Survey of Collaborative Filtering Techniques," Hindawi Publishing Corporation, Advances in Artificial Intelligence, Volume 2009, Article ID 421425, doi:10.1155/2009/421425, 2009, pp.1-9.

[28] Xiwang Yang, Harald Steck, Yang Guo, and Yong Liu, "On top-k recommendation using social networks," ACM RecSys'12 - Proceedings of the sixth ACM conference on Recommender systems, 2012, pp.67-74.

Applying Chatbots to the Internet of Things: Opportunities and Architectural Elements

Rohan Kar[1]
Hyderabad,
India

Rishin Haldar[2]
School of Computing Sciences and Engineering,
VIT University,
Vellore, India

***Abstract*—Internet of Things (IoT) is emerging as a significant technology in shaping the future by connecting physical devices or things with the web. It also presents various opportunities for the intersection of other technological trends which can allow it to become even more intelligent and efficient. In this paper, we focus our attention on the integration of Intelligent Conversational Software Agents or Chatbots with IoT. Prior literature has covered various applications, features, underlying technologies and known challenges of IoT. On the other hand, Chatbots are a relatively new concept, being widely adopted due to significant progress in the development of platforms and frameworks. The novelty of this paper lies in the specific integration of Chatbots in the IoT scenario. We analyzed the shortcomings of existing IoT systems and put forward ways to tackle them by incorporating chatbots. A general architecture is proposed for implementing such a system, as well as platforms and frameworks – both commercial and open source – which allow for the implementation of such systems. Identification of the newer challenges and possible future research directions with this new integration have also been addressed.**

***Keywords*—Internet of Things; Chatbots; Human-Computer Interaction; Conversational User Interfaces; Software Agents**

I. INTRODUCTION

The Internet of Things (IoT) is not just a well-recognized phenomenon but one that is shaping the digital age. It introduces an era of interconnected smart objects or 'things' developed upon existing Internet architectures. By using unique addressing schemes and standard communication protocols, IoT interconnects these things or objects thereby creating a varied range of technologies that can interact with each other and reach common goals [1].

An essential goal of connecting various sensors, actuators and services and processing data from them is to generate situational awareness and enable machines and human users to make sense of themselves and their surrounding environments.

The proliferation of IoT can be seen through the adoption of these "smart devices" in our daily life which include applications in Manufacturing, Agriculture, Medical and Healthcare, Transportation, Building and Home Automation and Energy Management among others. A report by Gartner estimates that there will be over 20 Billion connected things in activity by 2020 with Cisco estimating the number to be over 50 Billion [2, 3]. Among them more than half of all IoT endpoints in the consumer space alone. Hence IoT is a phenomenon which is certain to play a major role in our daily interaction with the digitally connected world.

A. Scope of Internet of Things

Literature presents various ways to define the Internet of Things. The RFID group defines Internet of Things as "world-wide network of interconnected objects uniquely addressable, based on standard communication protocols." ITU [4] defines it as "a global infrastructure for the information society, enabling advanced services by interconnecting (physical and virtual) things based on existing and evolving interoperable information and communication technologies."

While considering the broad vision of IoT, this paper focuses on the perspective of connected things and its applications. To do this, we simply create a separation of concern between the fragmented lower Open System Interconnection (OSI) layers of IoT and the unified adopted upper layers of IoT communication which use the World Wide Web and its standard network protocols.

The entire IoT system consists of Sensors (such as temperature, light, and motion), Actuators (such as displays, sound and motors), Computation (programs and logic), and Communication interfaces (wired or wireless). However, based on established advantages presented in prior literature [5, 6, 7, 8], our scope will be limited to interaction with IoT through Web Application Programming Interfaces (API) and in particular Hypertext Transfer Protocol (HTTP) based Representational State Transfer (REST) Architectures. A popular approach to Web of Things has been illustrated in Fig 1 based on [6].

The Evans Data Corporation (EDC) Report: Internet of Things - Vertical Research Service study [9] reveals that more than half of IoT developers connect to devices primarily through the cloud. The massive growth and acceptance of these cloud-based platforms such as IBM IoT Platform, Amazon Web Services IoT, Microsoft Azure IoT and Cisco IoT are indicative of its popularity among IoT companies. Hence, this paper also proposes the use of IoT cloud-based platforms in the proposed system design. This is discussed further in Section 4.

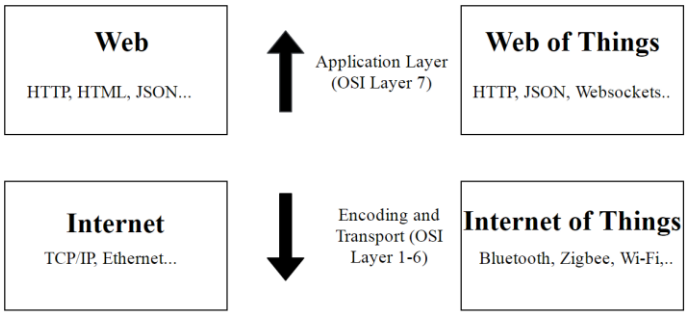


Fig. 1. Web of Things as shown in [6]

B. Scope of Chatbots

This paper proposes the use of Intelligent Conversational Agents. We refer to these as simply Chatbots (also known as Chatterbots or bots in general). Interestingly, there are many definitions for Chatbots in close relation with Software Agents (SA), Virtual Agents (VA) or Intelligent Personal Assistants (IPA) in literature and these have often been used in conjunction with each other. The term "Agents" itself has many definitions but among the earliest and most well-known uses of the term is [10] - "A self-contained, interactive and concurrently-executing object, possessing internal state and communication capability."

Software Agents can be most closely associated with Chatbots and has been well documented in prior literature [11]. The following key properties have been related to Software Agents [12]: (1) reactive, (2) proactive and goal-oriented, (3) deliberative (4) continual (5) adaptive (6) communicative, and (7) mobile. However, the purpose of this paper is not to explore the various types of Software Agents and agent-based systems or its properties but rather propose the solution to challenges faced in IoT through the use of the umbrella term for these Intelligent Conversational Agents, Software Agents or Chatbots as we refer to them. It is also important to note that Software Agent distinguishes itself from Intelligent Agents (also known as Rational Agents). Intelligent agents are not only computer programs. They can also be machines, humans or anything that is capable of a goal directed behavior [13].

Typically Chatbots are classified into two types: (1) Chatbots that function based on Rules (2) Chatbots that function based on Artificial Intelligence (AI). Chatbots that function on rules are often limited as they are only as smart as they are programmed. On the other hand, AI based Chatbots give the impression of being "intelligent" as they are capable of understanding natural language, not just pre-defined commands and get smarter as they interact more due to their ability to maintain states. Based on this, concepts such as Virtual Agents and Intelligent Personal Assistants (IPA) have come up, which use natural language processing, as well as speech recognition techniques. For example, Apple Siri, Amazon Alexa, Microsoft Cortana and Google Assistant are some of the popular IPAs.

In this paper, we present a novel paradigm combining these two disparate concepts of IoT and Software Agents in a single solution. However, the studies of these paradigms have largely been separate endeavors. We discuss how using chatbots as intelligent conversational interfaces can be used to address critical problems in IoT. We also propose a high-level conceptual architecture and discuss key architectural elements involved in communicating with an IoT system through Chatbots. To explain in the context of real-world applicability, we put forth existing solutions to each of the components in the architecture including frameworks, platforms and specify open-source tools which can be used to build such a system.

The remaining paper is organized as follows: In section II we discuss our motivation for introducing this novel concept of Chatbots in Internet of Things and discuss other literature work that has helped shape this idea. In Section III, we evaluate and examine the shortcomings and challenges of current IoT systems and the opportunity for chatbots to address them. Section IV proposes a system design and the key architectural elements. Finally, we present our concluding remarks by assessing opportunities and scope for future research and development in Section V.

II. Motivation and Background Study

The key to the massive adoption and diffusion of IoT is the proliferation of Internet in our daily lives. We use the internet to search for information, check emails, consume media, and connect with people via social networks and so much more. With around 40% of the global population (3.4 Billion) currently using the world wide web, this number is estimated to increase to 7.6 billion global internet users in 2020, a majority of which use mobile devices (such as phones, tablets, and wearables) [14]. Hence the internet has played a vital role as a global backbone for information sharing and interconnection of physical objects with computing/networking capabilities for applications and services spanning numerous use cases. The Internet alone, however, cannot address all issues of IoT. First, we will briefly discuss the challenges in IoT, and then mention the motivation for choosing intelligent conversational interfaces.

A. Challenges in IoT

Despite the wide scale efforts to popularize IoT, it still offers many practical challenges. Primarily, IoT systems operate in isolated technology or vendor specific silos which inhibit capability, value, and interoperability and create a widely disparate area [15]. Specifically, by restricting heterogeneous devices (such as home appliances, mobiles, tablets, embedded devices), sensors and services to communicate with each other across interconnected networks, possibilities of countless applications are hindered.

Secondly, the sheer number of actively connected things has already started to create problems in application, device and data management in IoT [16]. To address this issue, IoT platforms (such as IBM Watson IoT, Microsoft Azure IoT, AWS IoT) offer scalable, distributed cloud-based services to allow businesses to connect to an established infrastructure service or software quickly, without being concerned about backend complexities. While IoT Cloud platforms are a step in the right direction, offering many advantages, it still presents many challenges particularly in interoperability which has led to the issues of platform fragmentation [17, 18].

IoT systems also face a challenge of unifying User Interfaces (UI). It becomes increasingly difficult for users to keep track and access multiple applications, dashboards for every new "thing" in their ecosystem [19]. Hence unifying interfaces across multiple connected things and providing them with a high degree of smartness for improved user experience is a key challenge.

B. Relevance of Chatbots

According to a Business Insider study, Instant Messaging (IM) platforms (Such as Facebook Messenger, Slack, WhatsApp, and Telegram) have more active users than any other internet application including social networks, mailing applications etc. The same report shows that the top ten messaging platforms alone account for nearly 4 Billion users [20]. The global acceptance of chat based interfaces allows for ease of adoption and diffusion of newer technologies (such as Chatbot Applications) to be built on top of the pre-existing platforms. Therefore the global proliferation of chat as a Conversational User Interface (CUI) only furthers the motivation to develop interesting applications and use cases with chatbots.

Secondly, advancements made in the areas of Artificial Intelligence (AI), especially Natural Language Processing (NLP) have furthered the efficiency and quality of Chatbots allowing the user to make complex requests through simple natural language.

Finally, RESTful API's have been an important factor in the adoption of both IoT and Chatbots. The rationale behind it has been based on some of the following observations:

1) Ease of Development: Developers can take an API or service-oriented approach to development for both IoT as well as Chatbots. This means that application development methodologies would be the same with both embedded devices as with any web service (including Chatbots) that use Web APIs and in particular using RESTful architectures.

2) Ease of Deployment: Chatbot applications just like IoT applications can be designed and deployed on cloud platforms where developers need not be concerned about the underlying technologies such as Network, Storage, and Processing.

3) Standardized Web protocols: Owing to HTTP RESTful standards and protocols, it becomes technologically feasible and straightforward to integrate chatbot applications into IoT systems using application layer as the only concerned medium.

This ease of integration is a key motivation to develop platforms and frameworks which can synchronize chatbot applications within IoT platforms and frameworks.

Fig. 2. A Sample User-Chatbot conversation

III. Opportunities for Chatbots in IoT

The shortcomings of modern IoT systems can be broadly classified into two types: (1) Technology Centric Challenges and (2) Human Centric Challenges. With the help of example chatbot-user conversations given in Fig. 2, we discuss the opportunities for Chatbots and demonstrate ways in which Chatbots can overcome challenges in IoT.

Use Case (A)
User: "Make the living room temperature comfortable."
Chatbot: "Since the weather outside is 17 degrees Celsius,
I am setting the living room temperature to 21.4 degree Celsius."
Use Case (B)
User: "How far is my car charged?"
Chatbot: "The Tesla Model S is currently 40% charged. 3 Hours 10 minutes to full charge."
Use Case (C)
User: "Which lights are on?"
Chatbot: "The Guest Bedroom and Living room lights are ON."
User: "turn them off"
Chatbot: "The Guest Bedroom and Living room lights are now OFF."
Use Case (D)
User: "Help me setup my new device."
Chatbot: "Which device would you like to configure?
1) Smart Lock 2) Smart Kettle 3) Smart light?"
User: 1
Chatbot: "Ok, Enter your secret passcode for the smart lock."
User: "*****"
Chatbot: "Done. Smart Lock is now setup and ready for use."
Use Case (E)
Chatbot: "The monitoring service indicates that the smart lock has been offline for over 24 hours."
Chatbot: "Would you like me to report the issue to the Smart Lock Customer Support?"
User: "No, I want to talk to a human."
Human-Operator: "I can see the issue you are facing. I will try to resolve it remotely."

A. Technology Centric Challenges of IoT

1) Data Management: A key challenge in the realm of IoT is managing the vast amount of big data being generated, as IoT sensors are becoming easily affordable. Not only is the data produced by the sensors large but also diverse (varying in quality and type) and multimodal (e.g., temperature, light, sound, video) in nature. While data deluge is one challenge, drawing insights from the data and being able to present it in a timely, understandable way is a much larger challenge.

The well-known Knowledge Hierarchy also called the DIKW (Data, Information, Knowledge, and Wisdom) Pyramid can best illustrate the situation in the context of IoT [21, 22]. As one moves up the pyramid, the data gets smaller but becomes harder to gain abstractions and perceptions (Knowledge), which is required to derive actionable intelligence (Wisdom). Hence Chatbots are attempting to solve the problems of data and information management by mainly addressing the upper layers of the DIKW pyramid.

a) Data Context: Processing and analyzing of IoT data today is often done through the many "big data" solutions in cloud platforms which offer storage and computing infrastructure to accomplish the task. These existing IoT cloud solutions are capable of handling various data source and transmission challenges. However, a major challenge of existing IoT systems is conveying data about the different interconnected devices (sensors and objects) back to the user in a simple human understandable way. This requires context, which can be achieved by enabling Chatbots to understand the true intent of the user query and process information from their environments. Moreover, Chatbots have access to a global network of information via the internet and can be easily programmed to retrieve information in real-time which can improve the context.

In practical terms, Chatbots simplify the way we consume information from multiple screens and heavy data and graphics to simple Conversational User Interfaces (CUI) capable of delivering highly contextual and intelligible information within the flow of the chat app itself. Achieving this high-level of abstraction can deliver actionable intelligence (wisdom) with domain and user knowledge to maximize the full potential of IoT. For example, in use case (A), the user utterance was relatively vague. However, the Chatbot could have used contextual information from Real-time temperature along with knowledge of historical user preferences to perform a specific action.

b) Information Retrieval: Modern IoT dashboards are often saturated with various metrics, data points, charts and tables making it difficult for users to find the required information. Chatbots can effectively solve this problem by responding quickly to direct queries with highly accurate information. By understanding the specific intent of the user, they limit the scope of information to be presented. In terms of the knowledge hierarchy, Chatbots perform lookup and abstraction on IoT data. For example, in use case (B) the query only asked for Battery Charge related information. The Chatbot performed a lookup and limited the response accordingly.

2) Device and Application Management: A fundamental challenge of IoT has been the fragmentation of technology [17, 18]. Having application interoperability between heterogeneous devices from a single remote (mobile device or operation terminal) is especially uncommon. For example, a smart light and a Heating Ventilation and Air Conditioning (HVAC) system may belong to the same network and environment yet have different user control terminals which are mutually independent entities, unaware of each other nor able to communicate with each other.

Chatbots are built on IM platforms (such as Facebook Messenger and Slack) which support multiple different chatbot applications within itself. A single chatbot application is also capable of communicating with multiple IoT devices through unique HTTP REST APIs. Chatbots can thus act as a single interface for communication between single purpose devices (e.g., Controlling two smart lights), heterogeneous devices (e.g., Controlling an HVAC and a Smart Car) and even different IoT ecosystems (e.g., Controlling Smart home devices and Smart Retail devices) in the case of cloud-based IoT. For example, in the use cases above, the same chatbot is utilized to converse with multiple heterogeneous devices. Given the right permissions are available, it can even communicate with Public IoT devices.

3) Bridging Data across Platforms and Services: IoT platforms can be seen as software development environments which handle Device Management, Application Management, Connection Management, Dashboard, and Analytics. Owing to platform fragmentation [17, 18], sharing of data across platforms is still uncommon. One solution is to solve the issue at the application level by using third party services, which through APIs, can access data from each platform. For example, device data from an IBM IoT platform can be collected and processed by an analytics service along with sensor data from an Azure IoT platform thereby bridging the two data sources.

4) Search and Discoverability: A key attribute of IoT is the natural tendency of objects to be dispersed in the environment while being interconnected and identifiable at class-level (i.e. common information across the same class) or serial-level (i.e. unique to an individual object) [23].

Based on the permissions of the requester and the availability of the connected objects in the scope of the environment, IoT requires lookup and discovery services to find and control these objects effectively. Such services include the availability of sensors and actuators which the Chatbot would be able to retrieve from the entities and convey to the user at the appropriate times. In use case (C), the Chatbot was effectively able to find the active smart lights in the environment.

5) Monitoring and Reporting: From IoT wearables such as health monitoring devices to industrial sensors which convey information in real time, monitoring and reporting are the main aspects of IoT systems.

Chatbots can also be effectively used as monitoring services by integrating with solutions such as Application

Performance Management (APM). Accessing data from various IoT systems is a key advantage which is unique to Chatbots in this scenario.

Similarly, Chatbot services can utilize its reporting services and present the abstracted information to the user in an actionable and timely manner. In use case (E), the chatbot was monitoring the availability of the smart lock.

B. Human Centric Challenges of IoT

Chatbots were created with the primary purpose of improving the human-computer user experience. As such, solving the user experience shortcomings of IoT systems can be a significant opportunity for chatbots. IoT, with its complex system of applications, sensors, actuators and services present a daunting challenge of gaining technical knowledge to interact with these various components. Hence exposing settings and configurations to users presents an obvious and unfriendly burden that is far from ideal.

1) Cognitive Burden: The technology landscape of IoT is quickly changing. As newer features and use cases are introduced, there is an added responsibility to educate the end users which can be burdensome for both the users and the developers of the system. Complicated systems cause difficulties in adoption and diffusion. As an assistive technology, chatbots can simplify the learning curve by the following ways:

a) Help Texts: IoT-enabled Chatbots can feature help texts which clarify the user request to ensure that the action performed is same as the one intended.

b) Feature Recommendation: Chatbots can recommend possible actions to the user which can be made more intelligent and context-aware depending on user preferences and the dynamics of the environment.

c) Automating Tasks: Chatbots are good at automating common cyclic, tasks and can perform certain actions such as monitoring availability of sensors (uptime, downtime) and others through routine API calls, Websockets or Publisher-Subscriber methods.

d) Frequently Asked Questions: Feedback loops can be easily integrated within chatbots to aggregate most common queries, and this data can be used to improve the future Quality of Service(QoS).

As more use cases are discovered, chatbots can make the adoption and diffusion of IoT systems significantly easier and reduce the cognitive burden required to understand the functionalities of these systems.

2) User Interface Opportunities: Graphical User Interfaces (GUI) for IoT are largely functional in nature. While it achieves simplicity by displaying virtual switches, sliders, and buttons, it still has some shortcomings which Chat interfaces can solve: (1) Chat interfaces understand natural language which makes interaction with the system as simple as asking queries and receiving answers. There is no need for navigation of menus and finding the right icon/button to perform a task. (2) Chatbots use machine learning techniques to understand an individual user and can personalize the service to that user. In this way, chatbots can maintain the natural flow of the conversation as well. (3) They are also highly contextual interfaces and can understand the intent in the scope of the past interactions which is a unique feature of chatbots and speech-based systems. (4) CUI concern mostly textual information, therefore, simple log files can be maintained and consequently analyzed to make debugging easier.

3) Configuration Challenges: Apart from the knowledge required to adapt to the new systems and ease the cognitive burden, each IoT device has its unique setup and configuration in terms of software, network, firmware etc. As the number of different IoT devices increase, it becomes challenging and burdensome at best to navigate the interfaces of various applications and appropriately configure the system. Often technicians are involved in setting up and explaining the uses of the system.

Chatbots can guide and advise users on the right configurations for their system by creating step-by-step setup processes. This also reduces human effort involved in setting up the system. For example, a new device was configured in use case (D).

4) Lack of Automated Error Reporting: The distributed nature of most IoT systems implies that user report databases of IoT errors are spread across multiple organizations, Operating System (OS) vendors, Internet Service Providers, and device vendors which makes automated problem reporting a major challenge. Furthermore, users themselves are uncertain which organization to report the particular issue. Thus, various stakeholders in the system have a limited understanding of the actual nature of the problem and avoid sharing information with each other. Chatbots, in this scenario, can access these reported problems and by integrating other services, be able to not only retrieve information from the IoT system but send information to it. In Use Case (E), the chatbot identified the correct stakeholder to send the error.

5) Support Challenges: Remedying hardware and software issues in modern consumer IoT systems can be an irksome task. The recourse is to call the service provider for technical support or in many cases return the product. Either way, it is an unnecessary burden on the user as well as the support vendors in today's cost structure.

Smart Chatbots often have support services built into their functionality. Human-in-the-loop processes can be used to handle situations the Chatbot is not trained or authorized to perform, in real-time. In this manner, users need not go beyond the scope of the chatbot application to look for product support. Any software issue or hardware malfunction can be monitored, and Over the Air (OTA) software repairs can be performed. Chatbots can also be used to schedule technical repairs making it a convenient and fast solution to customer support [24]. In use case (E), a human operator was made to intervene.

IV. System Design and Architectural Elements

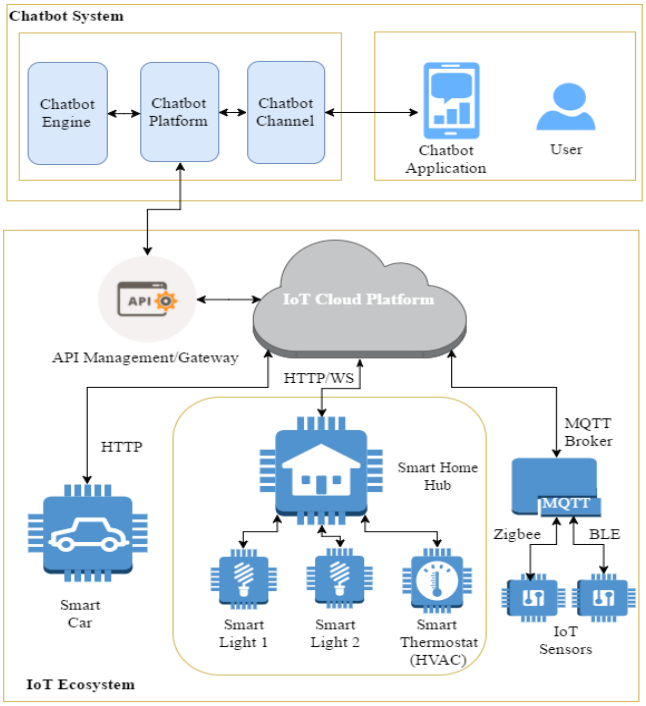


Fig. 3. Proposed System Design of IoT-Chatbot System

We present a conceptual system design which will aid in building Chatbot systems for IoT. Fig. 3 presents the high-level view of the overall architecture consisting of the IoT system and the Chatbot system.

A. IoT System

1) IoT Devices: In the context of the paper, we refer to an IoT device in the broad scope of the term as a "uniquely identifiable IoT endpoint which can be accessed and controlled using RESTful Web APIs." In the situation an embedded device does not have APIs, there are existing solutions to create APIs for them easily. For Example, Using platforms such as Zetta, one can create cloud-based IoT systems with full-fledged API support. In this paper, we consider a Home automation system consisting of Smart lights (e.g., Philips Hue), Smart HVAC (comprised of a smart thermostat e.g., Nest) as well as a Connected Car or Smart Car (e.g., Tesla). However, in principle, any IoT device may be considered for interfacing with Chatbots.

2) IoT Cloud Platform: IoT Cloud based Platforms is an important enabling technology in many IoT systems today. They deal with various fragmented technologies in embedded devices from access protocols (e.g., Message Queuing Telemetry Transport (MQTT), Bluetooth Low Energy (BLE), HTTP etc.) to different services, Software Development Kits (SDK), and integrations. There have been positive reports on the established advantages of Cloud-based IoT platforms [25].

Our system design stresses on accessing and controlling the embedded devices in question, such as Smart Car, Light, Thermostat etc., through the API Management/Gateway of the IoT Cloud Platform, regardless of the standards and protocols of the individual embedded devices as seen through the representation of IoT sensors (which use Zigbee and BLE) connected to a MQTT broker which interfaces with the Cloud Platform. Similarly, the Smart home hub and the Smart Car use different protocols to interact with the IoT cloud platform with HTTP requests or Websockets. Popular IoT cloud platforms today include Microsoft Azure IoT, IBM IoT, APIGEE IoT and Cisco IoT.

B. Chatbot System

1) Chatbot Channels and Platforms: Chatbot Channels are applications which run Chatbots on supported Mobile devices (e.g., Smartphones, Tablets, smartwatches) or Terminals (e.g., Desktop Applications).They are typically built on top of the existing instant messaging platforms. Popular Chatbot channels include Facebook Messenger, Slack, Telegram, Kik, Skype, Line and Twilio SMS. These channels are essentially the Chatbot applications in which a user interacts with the bot. The area of Chatbot development is still in its infancy, and there can be many different architectural approaches in implementing Chatbots.

In some approaches, the channels are interfaced separately with the Chatbot Platforms through connectors. In Fig 3, the Chatbot Platforms are hosted on cloud services which may use Webhooks to communicate with the Channel. It is important to note in this scenario that we consider text-based Input/output (I/O) of Chatbots to IoT. However, by using SDKs it is possible to integrate IoT to voice/speech-based commands as seen in Intelligent Personal Assistants such as Amazon Echo (which uses Alexa SDK) and Google Home (which uses Google Assistant SDK).

2) Chatbot Engine: Perhaps the most important component of a Chatbot is the engine, often referred to as Natural Language Understanding (NLU) engine. It is responsible for translating natural language into machine understandable action. Chatbot engines are often highly complex, using various NLP models and ML techniques to provide acceptable levels of accuracy. To make it easier for Chatbot developers, many companies offer the processing capability of the Chatbot engine as a Software-as-a-Service(SaaS) or 'AI-as-a-service' which are applied to Chatbot applications using APIs. For example, Wit.ai and Microsoft LUIS.

This paper is primarily focused on listing the relevant key components of the engine and its functionality in the context of IoT, not on designing the NLP techniques for the Chatbot engine. Next, we include key concepts typically associated with chatbot engines [26, 27]:

- **Entity Recognition**: Entities are domain specific information extracted from the utterance that maps the natural language phrases to their canonical phrases to understand the intent. They help in identifying the parameters which are required to take a specific action. To train the chatbot engine, entities which are expected to give the same actions are typically grouped together. Common entities can be predefined as they can be used in many different scenarios. For example,

Currency, Color, Date time, Location, Number etc. Domain specific entities can be trained to recognize similar phrases. IoT devices are one such domain specific entity.

For example, for the utterance: “Thermostat”, the acceptable phrases may be trained as "Thermostat", "heat", “heating”, “AC” ,"air conditioning" and will be decoded as {“type”:iot, “device”: “Thermostat”} where the entity is IoT. Entities may also have its own attributes. Example 2: The utterance- “$15” can be decoded as {“type”: “money”, “amount”:15, “currency”: “dollars”} where the entity is Money represented in JSON format.

- **Context Determination:** Determining the Context of the current user expression is an important feature of modern Chatbots. Understanding context can be important to handle situations where the utterances may be vague and have multiple meanings depending upon the history of the conversation. Contexts represent the ability of agents to maintain state (also called lifespan or the number of utterances after which the context will be removed) and match the requisite intent.

For example, if the user asks “Turn on the guest bedroom lights” asking about the bedroom room (location) in the first utterance and then in the following utterances, asks a typically vague statement such as: “turn it off” the Chatbot uses context to understand the second query relates to the earlier guest bedroom lights.

- **Intent Extraction:** Intents are the crux of conversational UI in chatbots. The intents represent what the users are looking to accomplish: get status updates, turn on/off devices, ask for help etc. The message passed from the user (utterance) in natural language is first analyzed for the intent. This implies, mapping a phrase to a specific action that should be taken by the IoT system as well as the specific dialog to be returned from the Chatbot. The information contained in an intent would be the context and action.

- **Action Classification:** Action refers to the steps that the IoT device will take when the intent of the user input is recognized. Actions have specified parameters which categorize details about it and triggered only when the intent recognizes them. For a smart home, the actions may be smartHome.lightsOn, smartHome.doorLock, smartHome.getStatus. In this scenario, other parameters may also be defined such as location (e.g., Dining room), time start/end (e.g., 10am, Thursday etc.), schedule (e.g., every hour, every minute) etc.

Once the action has been set and the minimum required parameters have also been defined, the right intent can be mapped to an IoT API endpoint. Meanwhile, the chatbot is required to maintain the natural flow of the conversation with the user. This is typically handled by a conversation module which models the semantics of the conversation. The quality of this conversation module varies from one Chatbot engine service to another, depending on the complexity of the AI. Hence in this paper, we assume the bare minimum case of returning a simple, relevant response to the user without going into the conversation module.

Chatbots can also be built using existing frameworks simplify the end-to-end process of creating and integrating Chatbots into messaging and IoT platforms. The key advantages of using a Chatbot framework are: (1) Ease of Development from pre-defined actions, integrations and SDK support for various IoT systems (2) Ability to ‘write once deploy anywhere’ through integrations with multiple Chatbot Channels. (3) Ability to use AI-as-a-service. For example, LUIS.ai in the case of Microsoft Bot Framework. Other popular Chatbot frameworks include API.ai and IBM Watson Conversation Service

V. Future Research Areas and Conclusion

IoT is poised to become intrinsic to the day to day activities in the future. However, the dynamic nature of IoT has its share of difficulties, and this paper has put forward the concept of using Chatbots to address some of the challenges in IoT. Through this initial endeavor we can identify possible areas that can be worked on for the future which show great potential:

1) Stronger AI-powered Bots: As more advances are made in the field of AI, Software Agents will also grow to become more intelligent in the future. The goal of Strong AI has been to match the machine’s intellectual capability to a human being. Immediate research challenges include improving decision-making ability to create more autonomous Chatbots. Better NLP as well as Natural Language Generation models will create more natural flows of conversation between humans and bots.

Most evidently, Chatbots will be able understand the opportunities in the conversation through improved context determination by seeking out knowledge from not only past history but a variety of external sources. Chatbots will also play a major role in the research areas of Intelligent Agents as well as Machine-to-Machine (M2M) research in IoT.

2) Cyber-Physical Systems and IoT: Cyber-Physical Systems refer to more advanced, next generation embedded Information and Communications Technology (ICT) systems. They share many similarities with IoT but with a higher combination and coordination of physical and computational elements [28]. The US National Science Foundation (NSF) identified cyber-physical systems among the key research areas in the foreseeable future [29]. IoT will play a major role in the transition to CPS as one of the key enabling technologies [16, 30]. Further advancements in AI aspects of Chatbots in IoT will be closely related to the Conversion, Cyber, Cognition and Configuration levels of the 5C CPS Architecture [31].

3) Wisdom of Things: The progress of Chatbots in IoT introduces the paradigm of feedback systems which has exciting research challenges in the areas of Wisdom of Crowds. The concept of wisdom of crowd suggests that aggregation of information can result in decisions that are

better than what could have been achieved by any individual in the group [32]. In the context of IoT, the sharing of big data from billions of sensors and devices creates more value in the ecosystem as compared to not sharing. However, it requires data interoperability rather than simply accumulating multiple disparate data sources which are incompatible or have no similarities. Hence Chatbots in IoT systems can use techniques such as Human Swarming, an approach that uses real-time feedback loops from groups of users to make accurate insights. There are plenty of interesting research opportunities in acquiring accurate values from the crowd.

4) Evolution of the Semantic Web: As the Internet itself changes, there are many more opportunities for exciting research in the areas of IoT and Software Agents. The development of a Web 3.0 or Semantic Web and its impact on the future of Software Agents has been clearly described in the literature [33]. This presents a great opportunity for research in IoT regarding Semantic interoperability which can have a major impact on the IoT paradigm itself. The evolution of the Semantic Web will have a major impact on areas in pervasive computing, M2M technologies resulting in Software Agents being able to draw more value and achieve a higher level of wisdom than before.

Development in the field of IoT has been phenomenal in recent times. Similarly, Chatbot systems are also becoming more intelligent and sophisticated as the days progress. To the best of our knowledge, no work has been published detailing the specific integration of Chatbots to IoT. This paper has attempted to integrate these two fields together by enlisting the key architectural components required and envision possible ways to address some of the present challenges in IoT. We hope that this endeavor will lead to more intelligent, efficient and integrated IoT systems.

REFERENCES

[1] Atzori, L., Iera, A. and Morabito, G., 2010. The internet of things: A survey.*Computer networks*, *54*(15), pp.2787-2805 J. Clerk Maxwell, A Treatise on Electricity and Magnetism, 3rd ed., vol. 2. Oxford: Clarendon, 1892, pp.68–73.

[2] Evans, D. "The Internet of Things How the Next Evolution of the Internet is Changing Everything (April 2011)." (2012): 346-360.

[3] van der Meulen, R., 2015. Gartner Says 6.4 Billion Connected 'Things' Will Be in Use in 2016, Up 30 Percent From 2015. *Stamford, Conn*

[4] ITU-T Recommendation database", ITU, 2016. [Online] Available:http://handle.itu.int/11.1002/1000/11559

[5] Guinard, D., Trifa, V., Mattern, F., & Wilde, E. (2011). From the internet of things to the web of things: Resource-oriented architecture and best practices. In Architecting the Internet of Things (pp. 97-129). Springer Berlin Heidelberg.

[6] Guinard, Dominique; Vlad, Trifa (2015). *Building the Web of Things*. Manning. *ISBN 9781617292682*.

[7] Vermesan, O.,et al.., 2011. Internet of things strategic research roadmap. O. Vermesan, P. Friess, P. Guillemin, S. Gusmeroli, H. Sundmaeker, A. Bassi, et al., Internet of Things: Global Technological and Societal Trends, 1, pp.9-52

[8] Guinard, D., Ion, I. and Mayer, S., 2011, December. In search of an internet of things service architecture: REST or WS-*? A developers' perspective. International Conference on Mobile and Ubiquitous Systems: Computing, Networking, and Services (pp. 326-337). Springer Berlin Heidelberg

[9] Evansdata.com. (2016). Evans Data Corporation | Internet of Things – Vertical Research Service. Available at: http://www.evansdata.com/reports/viewRelease.php?reportID=38

[10] Hewitt, C., 1977. Viewing control structures as patterns of passing messages. Artificial intelligence, 8(3), pp.323-364

[11] Nwana, Hyacinth S. "Software agents: An overview." The knowledge engineering review 11, no. 03 (1996): 205-244

[12] Schermer, Bart Willem. Software agents, surveillance, and the right to privacy: a legislative framework for agent-enabled surveillance. Leiden University Press, 2007

[13] Russell, Stuart Jonathan, Peter Norvig, John F. Canny, Jitendra M. Malik, and Douglas D. Edwards. "Artificial intelligence: a modern approach". Vol. 2. Upper Saddle River: Prentice hall, 2003.

[14] Broadband Commission, 2014. The state of broadband 2014: Broadband for all. Geneva, Switzerland: The United Nations

[15] S. Liang, "SensorThings API - connecting IoT devices, their location and theirdata,"2016.Available:http://www.eclipse.org/community/eclipse_newsletter/2016/march/article2.php

[16] Miorandi, D., Sicari, S., De Pellegrini, F. and Chlamtac, I., 2012. Internet of things: Vision, applications and research challenges. Ad Hoc Networks, 10(7), pp.1497-1516

[17] Celesti, Antonio, Maria Fazio, Maurizio Giacobbe, Antonio Puliafito, and Massimo Villari. "Characterizing Cloud Federation in IoT." In 2016 30th International Conference on Advanced Information Networking and Applications Workshops (WAINA), pp. 93-98. IEEE, 2016

[18] M. Wallace, "Fragmentation is the enemy of the Internet of Things | Qualcomm",*Qualcomm*, 2016

[19] M. Littman and S. Kortchmar, "The path to A programmable world," 2014.Available:http://footnote1.com/the-path-to-a-programmable-world/

[20] W. Mckitterick, "The Messaging App Report: How instant Messaging can be monetized," Business Insider

[21] Rowley, Jennifer E. "The wisdom hierarchy: representations of the DIKW hierarchy." *Journal of information science* (2007)

[22] Barnaghi, P., Wang, W., Henson, C. and Taylor, K., 2012. Semantics for the Internet of Things: early progress and back to the future. International Journal on Semantic Web and Information Systems (IJSWIS), 8(1), pp.1-21

[23] Bandyopadhyay, Debasis, and Jaydip Sen. "Internet of things: Applications and challenges in technology and standardization." *Wireless Personal Communications* 58, no. 1 (2011): 49-69

[24] Vinyals, Oriol, and Quoc Le. "A neural conversational model." *arXiv preprint arXiv:1506.05869* (2015)

[25] Google, "Overview of Internet of things," Google Developers, 2016. Available: https://cloud.google.com/solutions/iot-overview

[26] Microsoft, "LUIS: Help," 2016. Available: https://www.luis.ai/Help

[27] API.ai,"Api.ai"2016. Available: https://docs.api.ai

[28] Rad, C.R., Hancu, O., Takacs, I.A. and Olteanu, G., 2015. Smart monitoring of potato crop: a cyber-physical system architecture model in the field of precision agriculture. Agriculture and Agricultural Science Procedia, 6, pp.73-79.

[29] Wolf, Wayne (November 2007). "The Good News and the Bad News (Embedded Computing Column".IEEE Computer. 40 (11): 104–105.doi:10.1109/MC.2007.404

[30] Wan, J., Chen, M., Xia, F., Li, D. and Zhou, K., 2013. From machine-to-machine communications towards cyber-physical systems. *Comput. Sci. Inf. Syst.*, *10*(3), pp.1105-1128

[31] Lee, J., Bagheri, B. and Kao, H.A., 2015. A cyber-physical systems architecture for industry 4.0-based manufacturing systems. *Manufacturing Letters*, *3*, pp.18-23

[32] Joe Barkai. (2016). Wisdom of Things - Joe Barkai. Available at: http://joebarkai.com/wisdom-of-things

[33] Berners-Lee, T., Hendler, J. and Lassila, O., 2001. The semantic web. Scientific american, 284(5), pp.28-37

State of the Art Exploration Systems for Linked Data: A Review

Karwan Jacksi
Computer Science Department
University of Zakho
Zakho, Iraq

Nazife Dimililer
Information Technology Department
Eastern Mediterranean University
Gazimagusa, N. Cyprus

Subhi R. M. Zeebaree
Department of IT, Akre Technical College
Duhok Polytechnic University
Akre, Iraq

***Abstract*—The ever-increasing amount of data available on the web is the result of the simplicity of sharing data over the current Web. To retrieve relevant information efficiently from this huge dataspace, a sophisticated search technology, which is further complicated due to the various data formats used, is crucial. Semantic Web (SW) technology has a prominent role in search engines to alleviate this issue by providing a way to understand the contextual meaning of data so as to retrieve relevant, high-quality results. An Exploratory Search System (ESS), is a featured data looking and search approach which helps searchers learn and explore their unclear topics and seeking goals through a set of actions. To retrieve high-quality retrievals for ESSs, Linked Open Data (LOD) is the optimal choice. In this paper, SW technology is reviewed, an overview of the search strategies is provided, and followed by a survey of the state of the art Linked Data Browsers (LDBs) and ESSs based on LOD. Finally, each of the LDBs and ESSs is compared with respect to several features such as algorithms, data presentations, and explanations.**

Keywords*—*Exploratory Search System; Linked Data; Linked Data Browser; Semantic Web

I. INTRODUCTION

Over the time, World Wide Web (WWW) has made data sharing online an easy task for all users which created an immense amount of data available and transformed the Web to a massive semi-structured database [1]. Hence, retrieving relevant information efficiently from this huge dataspace is a huge challenge. Usually, search engines are used to retrieve data from the current Web, however, to get accurate retrievals, very efficient indexing and seeking strategies and increasingly more complicated heuristics must be employed.

The search approaches have two main categories: lookup (or keyword-based) and exploratory search [2]. Lookup search approaches, typically, have database systems in the background and information is retrieved based on the keywords given. In this widely used search approach, the data is mainly textual documents, and the search items are known [3].

Exploratory search category is a specific information searching approach where the user targets and objectives are not necessarily known during the searching process. The users of this category concentrate on learning and investigation rather than fact retrievals and query answering. They compare, analyze and discover new concepts for the retrieved information [4].

Finding information on the existing Web, also known as the syntactic Web, is based on keyword search, which has limited recall and precision due to synonyms, homonyms…etc. of the keywords. Therefore, the quality of obtained results is rather poor. To enhance it, annotations are added to the contents of the syntactic Web forming the Semantic Web (SW) [1] [5].

The SW is an extension of the syntactic Web through standards by the W3C[1], where data is given well-defined meaning, can be understood by machines, thereby allowing machines and people to work in collaboration [6]. It uses W3C standards such as Resource Description Framework (RDF) to support unified data formats and exchange protocols over the Web so that the meanings of data are understood by machines. This technology increases the efficiency of search engines by enabling machine-driven data processing [5]. The RDF is known as the HTML of the SW and it is purely an XML language. It is the W3C standard used to represent information in the Web and describes Web resources as *<subject, predicate, object>* components, which is known as triples. The structure of interlinking resources of the Web is extended by RDF to use Uniform Resource Identifiers (URIs) to indicate the connections among resources forming a directed and labeled graph of structured and semi-structured data.

Ontologies are essential units of the SW infrastructure and more often known as the backbone of SW [7]. Web Ontology Language (OWL) and RDF Schema (RDFS) are the knowledge representation languages and data models recommended by W3C where fundamental elements for the description of ontologies are given[2].

In the SW, the Web contents enhanced with data annotation to be linked with each other forming the Web of data which makes it possible for relevant data to be found once only a subset is given. The terms SW and Linked Data (LD) have been coined by Berners-Lee and describes the LD as "the SW done right"[3] [8].

[1] World Wide Web Consortium: www.w3.org

[2] http://www.w3.org/RDF

[3] http://www.w3.org/DesignIssues/LinkedData.html

LD term indicates a couple of stages or rules to publish and link structured data on the Web. The stages have been identified by Lee in his notes about design issues on Web architecture and shortly became the principles of LD [9]. These stages are 1) things should be identified with URIs; 2) HTTP URIs should be used for these URIs so that people can reuse them; 3) use standards such as RDF and SPARQL[4] when providing information for users looking up URIs; 4) link to other URIs so that further things are discoverable [10].

In fact, LD is to publish data on the Web wherein the data is understandable and processable by machines, its meaning is defined clearly, and is linked to/from other external data sets. Typically, untyped hyperlinks are used to connect HTML data with one another in the Hypertext Web, whereas LD relies on RDF data so to build typed links to connect things worldwide forming Web of Data[5] [11]. When the LD is available under an open license, it is called Linked Open Data (LOD).

The main target of this work is to present the current state of the art LD based exploration systems and to compare them regarding their features and underlying technologies so as to identify their strengths and weaknesses. A broad purpose of our study is to show to what extent the subject of exploration systems for LD has arrived, what kind of algorithms and technologies are used, and which Linked Datasets are most utilized. This is of great importance for the researchers in this field to effortlessly find and compare the above points.

To the best of our knowledge, there is only one other survey study concerning to the topic of exploration systems based on LD presented by [12], where Linked Data Browsers (LDBs), Linked Data Recommenders (LDRs) and Exploratory Search Systems (ESSs) are to some extent reviewed. However, none of the LDBs examined in our paper are in the found study, since in this paper only LDBs from 2012 to this point of time are considered, while the LDBs studied in the found research are up to the year 2011. Conversely, LDRs are explained in the found paper but not in our paper.

This research is conducted by searching and collecting articles from leading sources which are either found from databases such as Web of Science (WoS) or Scopus, or from SW Conferences and Challenges. 28 systems were selected and reviewed in details, 16 systems were chosen to be included in this paper. The selection of papers to be included in this paper based mainly on the authoritative paper sources (publishers) and on the publication date (most recent systems are included).

This paper is organized as follows; Section 2 reviews each of the LDBs in details; Section 3 presents a comprehensive review for the existing LD based ESSs; extensive comparisons of the systems are discussed in section 4, and finally we draw some conclusion in Section 5.

II. LINKED DATA BROWSER (LDB)

In the recent years, the use of LOD has notably increased on the Web. Nevertheless, it remains challenging to be used by users, particularly lay-users. Since the SW foundation, the interaction with LD and its visualization have been documented as issues [13]. Since then, a plethora of LDBs allowing users to understand, explore and interact with the massive LOD have been developed. Some of which, such as Tabulator[6] and Explorator[7], present data as pairs in tables and others, such as Graphity[8] and RelFinder[9], are graph-based browsers while other approaches, such LODmilla, combine features from both. In this section, we will review some of the states of the art LDBs to give an overview of their functionalities and features.

A. LodLive

LodLive is an LDB which uses standards of the LD to navigate RDF resources with the aim of spreading the fundamentals of LD in dynamic visual graphs through a user-friendly interface [14]. Within the application, resources located in different endpoints can be linked so as to discover unexpected connections. Also, inverse relations can be navigated even for different endpoints.

LodLive, via Sesame Framework, is able to parse RDF data even when they do not reside in an SPARQL endpoint. This can be done by generating a graph remotely to store temporarily the requested resources for making queries. During the initial stages of ontology definition, LodLive is a useful tool to verify the validity of an RDF schema and visually pick a solution among several.

JavaScript application layer has been used in the system without the necessity of any application server by parsing the JSON formatted retrievals of JSONP (JSON with Padding) calls from endpoints, and presenting them in an HTML5 web page.

B. CubeViz

CubeViz is a faceted navigation browser and an extension of OntoWiki[10] tool to visualize statistical data represented in RDF [15]. Usually, statistical data sets known as data cubes or basically cubes are distributed as spreadsheets or bi-dimensional matrices. To extract triples from these spreadsheets, further tools such as *CSV2DataCube*[11] (a plug-in extension in OntoWiki) are needed. These tools, usually, use RDF Data Cube vocabulary[12] which is an *SDMX*[13] standard and state of the art in representing statistical data in RDF. Thus, tools distributing multidimensional statistics on the web often use RDF Data Cube vocabulary so that they are able to be linked to related RDF datasets.

CubeViz is built so as to conceal the complication of RDF Data Cube vocabulary for users and to encourage the navigation and investigation of cubes. It utilizes the RDF Data Cube vocabulary to produce a faceted browsing for statistical data that is able to visualize and filter interactive explanations

[4] http://www.w3.org/TR/rdf-sparql-query
[5] http://www.w3.org/TR/2004/REC-rdf-concepts-20040210
[6] http://www.w3.org/2005/ajar/tab
[7] http://www.tecweb.inf.puc-rio.br/explorator
[8] http://graphity.org
[9] www.visualdataweb.org/relfinder.php
[10] OntoWiki: a tool provides cooperative support in creating RDF knowledge bases and their maintenance and publication.
[11] https://github.com/AKSW/csvimport.ontowiki
[12] http://purl.org/linked-data/cube
[13] https://sdmx.org

in diverse graph types such as charts and scatter plots. The JavaScript library *Highcharts*[14] has been used for rendering visualized charts for client side users, while on the server side, a PHP class called *PHPlot*[15] is used to render charts.

To use CubeViz, a preferred SPARQL endpoint and data structure have to be chosen. Later, aggregated modules which are defined as instances have to be selected. These instances reference various types of module properties.

C. Linked Data Visualization Model (LDVM)

LDVM can be used to quickly make representations of RDF data visually [16]. It permits users to connect to and extract data from different datasets with different visualization techniques. The conceptual framework of the model is based on Data State Reference Model (*DSRM*) model offered by [17] as a result of adopting its stages' operations, names and transformations so as to fit the LD environment.

The LDVM model may well be seen as a pipeline and it consists of four stages to process data, and three transformation operations in between these stages. The stages are *RDF data:* the raw data, *Analytical extraction*: to extract data from a previous stage, *Visual abstraction*: the visualizable data and *View*: present the data in different views. While the operations in between these stages are: *Data transformation*, *Visualization transformation, and Visual mapping transformation*. So, the model can be sectioned into two core sections: data space and visual space.

To proof the LDVM concept, a prototype and a useful RDF data browser is built based on the LDVM model called *LODVisualization*[16]. It gives multiple visualizations while browsing RDF datasets. Any endpoint supporting JSON and SPARQL 1.1 is compatible with this prototype. Several visualizations, such as charts, tables, treemaps, can be shown.

The server side of *LODVisualization* is written in Python, whereas the client-side built using HTML, CSS, and JavaScript mostly using *D3.js* library[17] and *InfoVis Toolkit*[18]. Google App Engine (GAE)[19] is used as a cloud computing platform. To avoid duplicated queries when users choose different visualization views, a cache system using GAE *Datastore* and *Blobstore* is used on the server so the performance and scalability are improved.

D. LODmilla

LODmilla is a generic LDB to discover and edit LOD with the ability to combine the features of textual and graph based LOD browsers [18]. The web application provides viewing and searching LOD graphs as well as other browsing services such as editing and reorganization utilities for data located in various linked datasets. The system uses a dedicated server for its search processes so as to support caching and rapid triple loading.

To load triple data, the approach switches through two methods employed: the SPARQL-based query and actionable URIs. Through the Jena toolkit at the server side, RDF data is able to get parsed into several serializations including JSON. As a result, a wide range of LOD datasets is possible to be used concurrently despite the configurations of datasets at the client side, hence the graph data is fetched using the actionable URIs method.

Graph traversal techniques have been applied in the system to find paths between notes. It starts from both sides of nodes using basic heuristics excluding connections having huge connections so as to discover next level paths quickly.

LODmilla can be used as an LD editor since it provides utilities to add/remove nodes and/or draw/cut out edges between the nodes in the graph. This editing facility is useful for finding quick solutions for incomplete graphs.

E. SView

Smart View, or SView, is a system that allows users to navigate entity descriptions for LD perceptively [19]. It uses Lenses, a group of organized features, to cluster and organizes entity descriptions so as to support users find related information easily. Moreover, it provides different methods to discover related elements, such as discoveries depending on link pattern and discoveries based on entity similarity.

Through SView, registered users are able to create their personal lenses so that they personalize how the features of LD are grouped and ordered. Then, they are able to share the created lenses globally so that other users can use or customize them for their needs. Accordingly, the system is able to provide a set of already created lenses for all users through leveraging these personalized lenses of the registered users and offer them to other users globally.

In the system, if a value of the current entity feature is a new entity, this is called a link. While discovering an entity description over lenses, users follow the link to the entity to discover related entities. Along with this method, two other mechanisms for discovering related entities are offered by the system: link pattern discovery and similarity based entity recommendation.

In the link pattern discovery method, having more than one link from an entity to another entity is possible in LD. The set of all links from between them is called entity's link pattern. Linked entities having a shared link pattern are clustered for exploration. Consequently, descriptions of entities having shared link patterns can be navigated by users. As a result, the system is able to provide pivoting from an entity to linked set of entities by transforming lenses to tables where the rows in the table are entities and the columns are features.

In SView, similar entities can be recommended using several approaches. The system uses surfing history of entire users to recommend similar entities. One of the approaches is to discover entities having mutual types to the entity in use. Thus, more mutual types, more recommended entity.

[14] http://www.highcharts.com
[15] http://sourceforge.net/projects/phplot`
[16] http://lodvisualization.appspot.com
[17] http://d3js.org
[18] http://philogb.github.io/jit
[19] https://en.wikipedia.org/wiki/Google_App_Engine

F. LD Viewer

Reference [20] implemented an adjustable framework and LDB that is integrating several tools to generally provide an ease of use LD explorations. The framework objective is to deliver a unified interface with a set of powerful features so it can without difficulty be adopted by several linked datasets.

The important component of the system is the property table which presents all the obtainable properties of the explored entity. It proposes forward and reverses properties for each explored entity together with pagination facility for reverse properties that have a huge amount of values.

Labels are presented rather than resource URIs so as to get a better legibility of the property table. Live previews for each link is also available in the system, by showing a brief preview of links in the property table. OpenStreetMap API[20] is used to generate maps for entities having location information.

For each triple in the property table, there is a clickable action (or set of actions) related to it when integrated conditions are met. These actions differ from each other due to the nature of the triple. For instance, using one of the actions, the user can make annotations to DBpedia dataset if the action is applicable for such triple.

The architecture of the application takes advantages from Model View Controller (MVC) software architectural pattern, it has been built primarily with AngularJS[21] framework, and modules of *Jassa library* [22] (JAvascript Suite for Sparql Access) are reused. To adopt the interface for datasets, there is no need to understand the core layer of the system.

Theoretically, to set up the framework to a dataset, three levels have to be concerned: 1) the triple store, which is reachable via SPARQL query language; 2) the server side to let the JavaScript implementation to work; and 3) the client side, which is implemented purely in JavaScript.

G. DBpedia Mobile Explorer

DBpedia Mobile Explorer is a mobile application to explore and visualize LD [21]. The framework is implemented to operate as a generic and a domain specific explorer for DBpedia dataset. In the generic exploration case, concise details of a resource been explored are given, a graph is formed for the connected nodes to the resource, and a table is created to present all important materials. In addition, all of the related categories to the resource are fetched from the DBpedia and listed to the user so that they can explore these categories and find other resources contained within that category. The specific domain case drives users to only explore resources of a specific domain such as music or films. Therefore, it is infeasible to explore resources beyond the specified domain rather than presenting concise details of them.

MVC software pattern is considered in the framework. The model in MVC contains classes with reference to resources, and a parser to fetch the classes from the resources received by DBpedia. While controller and view, which are the primary units of the framework, have been designed independently of models so it can be reused with diverse modules.

For the domain specific exploration, RDF:type property is examined prior resource visualization so as to exclude out of domain resources and to only generate short details for them. The domain to be specified is not fixed and it can be specified by domain experienced users. The application is considered for DBpedia dataset, but it is able to be applied to other linked datasets as well as it depends on RDF and SPARQL.

H. DBpedia Atlas

DBpedia Atlas is a web application that allows users to browse classes, instances, and relationships of DBpedia dataset in an interactive and map-like visualization. The application is built by adopting the efforts of [22] on Gosper Treemaps to Linked datasets with a structure of the hierarchical ontology. Therefore, it transforms entities of the DBpedia dataset to be shown as maps creating a hierarchy of areas along with their ontological class. A group of thematic maps and accessory charts are then created on top of these maps creating an atlas to describe diverse features of the dataset. The aim of the project is to give inexpert SW users a way to explore the dataset and understand its fundamental features.

System interface includes three main modules: map, search box, and infobox modules. The complete instances and classes of DBpedia are delivered to users through the map. Hence, the DBpedia root, e.g. *owl:Thing*, forms the primary map and the connected entities to the root form the regions inside the map. A minor island (separated from the primary map) is also created for untyped instances. For each region, there is a label, but smaller regions have to be zoomed so as to make the labels visible. However, to give users a further identification of basic categories, a manual identification for some of the instances are set with permanently visible labels. When an instance is selected from the map, all its details are presented in the infobox, and all its linked instances are distributed by way of red dots on the map. Several thematic maps are also available within the application, for instance, the map can illustrate the depth of the classes in the ontology hierarchy by notions of the darker colors the deeper in the hierarchy [23].

III. Exploratory Search System (ESS)

ESS forms a special category of seeking information on the Web with the purpose of revealing related information to the searcher along with retrievals of what have been searched for. With this search category, the final targets of the search are not known, and the goal itself is not defined. Therefore a set of additional activities, for instance, learning, exploration and evaluation, are accessible through this category [24]. The history of this category has begun with the Exploratory Search Interface XSI 2005 Workshop [25]. The machine data processing nature of Semantic Web (SW) proposes overwhelming possibilities for search engines specifically for ESS [26]. In this section, a review of the existing ESSs based on the SW technologies is presented. A more detailed survey is presented by [27].

[20] http://wiki.openstreetmap.org/wiki/API
[21] https://angularjs.org
[22] http://aksw.org/Projects/Jassa.html

A. Yovisto

The authors of this system address the issue of deploying explorative search for video data using SW technology and LD [28]. They confirm how LOD can be used to enable an ESS for video data. Yovisto is a search engine dedicated to academic lecture and conference videos. It provides an exploratory search property based on LD. Through Yovisto's time-dependent index, the ability to search in video contents is viable. This feature makes Yovisto different from other video search systems. Machine analysis methods e.g. smart character recognition and scene detection approaches are used to generate metadata. Additionally, annotations and time-based tags can be given by users within their comments. The index of Yovisto is generated through fine granular time-based metadata

The system utilizes LOD resources while providing search results; it presents additional materials that are relevant to searchers' query semantically. DBpedia dataset is used to provide the relevant information. Since DBpedia dataset is enormous and there is a huge amount of data for each entity, allocating all the related information for each entity is a very heavy process. Therefore, statistically based heuristics are created to select the best relevant retrievals for the searchers' query. Running online queries against DBpedia is a tedious task and consecutively falls in performance issues particularly if many related resources need to be returned. Therefore, an offline processing has been set up to process every term in advance.

B. Semantic Wonder Cloud (SWOC)

SWOC is an ESS that uses DBpedia as its dataset allowing surfers to navigate through the dataset using its semantic connections [3]. On top of utilizing the DBpedia semantic properties, the system takes advantage of external resources, like search systems and social tagging approaches, to find and present the DBpedia resources. So, this hybrid feature makes the system distinctive among other ESSs to rank the resources. The system consists of two main components: back-end, where the links between pairs of DBpedia resources are computed, and a front-end, where the attained data from the back-end is presented in a flash based interface.

For the ranking of the results of DBpedia resources, DBpediaRanker is used, which is an algorithm to compute the similarity of the resources compared to the initially queried resource. Thus, when the DBpedia resources are explored the DBpediaRanker uses external sources to compute and find the value of similarity for each pair of resources founded in the graph exploration.

In the front-end, the topic is selected, using the DBpedia lookup service[23], from a drop-down list of resource labels. The selected resource is returned together with its 10 most related resources forming a star network topology like a graph. The central node is the selected resource, while the surrounding nodes are the most related resources to the central one. The surrounding resources have different sizes depending on the similarity value; the bigger, the most similar to the central node. An info box to the right of the graph is presented showing a short description of the central node.

C. Lookup Explore Discover (LED)

LED is an ESS aiming to expand the search progress for users through providing an accurate exploration related to their queries [29]. DBpedia dataset is used in the application so that semantically related data are delivered. Users select the concept of interest from an autocomplete dropdown list of resources using DBpedia Lookup Service. Then, the system proposes a collection of concepts that are related to the user query forming a cloud of tags. Improving the results are possible through the system, by adding the proposed tags to the query bar so a combination of the two concepts is created and is sent as a new query to the system. Relevant information to each of the individual concepts and the combined concepts are returned and presented at separate tabs. As soon as a new tag is added to the query, a new search is performed and new tabs are added to the tab bar. All of the concepts in the tag cloud are associated with each other semantically.

LED is also presenting results of the user query from external search and microblogging systems, which can help users get more general results of their queries, or where the results cannot be found in the dataset so a broader search is given. DBpediaRanker is utilized by the system for ranking resources for the user query. A RESTful JSON API is provided by LED so that any system with HTTP requests can access the LED. This is useful for building web applications featuring new methods and it is suitable for various algorithm comparisons.

D. Aemoo

Aemoo provides exploratory search over the Web [30]. It exploits Encyclopedic Knowledge Pattern (EKP) to deliver its ESS. The system uses DBpedia and external sources to resolve user queries. The combined information is derived from Wikipedia, Twitter and Google News sources. Information combination is attained based on cognitively-sound principles by using knowledge patterns, hypertext link structure and SW technology.

Result presentation is based on EKP[24] filtration so that only relevant results from the returned information are presented. Additionally, the reason for presenting only those results is provided. The system proposes a further utility called *curiosity* so as to present additional knowledge which has been filtered by EKP.

E. Seevl

Seevl is an application for exploring musical data based on SW technology [31]. The application mines music connections to collect desired information with the aim of making context, search and discovery possible to be brought to users like music. A linked dataset of musical entities is built from collecting several resources on the Web. These entities contain musical data for instance Bands, Artists, and etc. so that services, such as recommendations, on top of the dataset are given.

[23] http://dbpedia.org/projects/dbpedia-lookup

[24] www.ontologydesignpatterns.org/ekp

A Virtuoso powered RDF store is created for the data being collected, and then it is hosted on an Elastic Compute Cloud (EC2) so as to get the EC2 architecture advantages like elastic cache and load balancing. Moreover, an LDB is built for this data so that it can be browsed by the user and also provides recommendations for the available artists.

F. Discovery Hub

Discovery Hub is an ESS based on the SW technology [32]. The system uses DBpedia data to fetch data from. DBpedia lookup service is used to permit users to select their topic of interest. While selecting the topic a stack is created so that new topics can be selected in case of finding relationships among them. The results of the query are ranked and categorized based on their similarity. Entity labels are used to rank and present retrievals, and explanations for how the results are presented are provided by the system in a graph form. The engine queries the DBpedia SPARQL endpoint on the fly without the need of preprocessing necessities.

Semantic spreading activation is extended on typed graphs of the LD formality and integrated with a graph sampling technique so the results are computed and returned to the frontend.

G. Linked Jazz

Linked Jazz is an ESS to reveal the relationships among jazz musicians based on Linked Jazz dataset [33]. It exploits the technology of LOD to improve the exploration of cultural heritage materials and to enhance the semantics describing them and to reveal the relationships among musicians and expose their society networks from the resources based on transcripts of interviews from jazz archives. Therefore, an RDF store describing these relationships is created as LOD.

The fundamental mechanism for the LOD to be built is that a unique ID (URI) has to be assigned to each entity in the directory. Thus, a directory of jazz musician's names combined with their URIs is constructed. The directory is called Linked Jazz Name Directory. Mapping tool, the dataset foundation application, ingested by extracted files from DBpedia and bibliographic name authority creating the Name Directory. Later on, this directory is refined by Curation tool. The process of Mapping and Curation tools is achieved by automated processing and crowdsourced activities.

The created dataset is then explored via Linked Jazz Network Visualization tool. The tool offers several visualization models such as Dynamic, Fixed, Free and Similar models so that users can use the appropriate model for their exploration of the jazz musician's network.

H. inWalk

inWalk is a Web application to explore LD based on inCloud and thematic walk ideas [34]. The concept of inClould refers to a high-level thematic graph where the vertices of the graph are clusters of relevant LD, and edges are associations of proximity amongst graph nodes. Hierarchical clustering algorithm HC^{f+} is used to construct inClould [35]. The application is built with targets to 1) defeat inflexible LDBs through a presentation of a thematic and a high-level data views generated from similarity based combination techniques; 2) provide common querying methods for inexperienced users in RDF query languages.

The system features can be summarized as 1) Abstraction by aggregation: by providing a conceptual view of data via inCloud notion; 2) exploration by walks: by providing the LD exploration through thematic and inside walks; 3) filtering by patterns: by enabling users to only explore a portion of inClould that satisfies them through filtering actions over the inCloud structure.

There are two main components of the system; the engine, which transforms retrievals from linked datasets, like Freebase and DBpedia, to a similar inCloud cluster; and the front-end, which is an HTML5 based interactive interface, to deliver discovery walks on inCloud clusters.

IV. Discussion

TABLE I presents a summary to LDBs reviewed in the previous section. From there, it is observable that most of the browsers have the ability to connect to more than a specific dataset and present their data. DBpedia has been used in all of them as a default dataset except for CubeViz where it depends on the datasets having statistical contents. The systems present the data in different visualization techniques. LodLive, LODmilla, and DBpedia Mobile Explorer present the retrieved data in graph forms. Infobox-like navigation has been used in LodLive, LODmilla, and DBpedia Atlas in addition to their primary presentation of data so as to explore detailed information about the selected node. CubeViz presents data in statistical forms such as different charts and scatterplots. Faceted navigation is offered by LodLive, CubeViz, and LODmilla so as to let users easily filter the retrieved data and discover their interest. All but CubeViz and LODVisualization use lookup as for query paradigm, while the remaining use dataset existing dimension elements and manual selection of entities respectively. Query suggestion feature is only proposed by LodLive browser, this feature gives users a new opportunity to figure out how the retrievals come from the datasets through SPARQL queries.

Four out of eight systems propose extra utilities so they can be used as LD editors, this can be seen in LodViz, CubeViz, LODmilla and LD Viewer.

The specifications of this feature vary from a system to another. However, this is a great feature for users to be available for both lay users and SW professionals. Lay users can benefit from this feature by obtaining a better understanding and learning for how LD are constructed, and how to add/edit new/existing graph properties or nodes, and for professionals to construct a new knowledge or make advanced practices in the dataset.

TABLE I. LDBS SUMMERY

System Name	LodLive	CubeViz	LODVisualization	LODmilla	SView	LD Viewer	DBpedia Mobile Explorer	DBpedia Atlas
Release Date	2012	2012	2013	2014	2014	2014	2015	2015
Web Address	http://en.lodlive.it	http://cubeviz.aksw.org	http://lodvisualization.appspot.com	http://munkapad.sztaki.hu/lodmilla	http://ws.nju.edu.cn/sview	http://ldv.dbpedia.org	Mobile app	http://wafi.iit.cnr.it/lod/dbpedia/atlas
Dataset(s)	Multiple	Any dataset contains statistics	Multiple	Multiple	DBpedia	Multiple	DBpedia	DBpedia
Visualization Technique	Graph & infobox	Charts & scatter plots	Treemap, tree, charts	Graph & infobox	Tables via Lenses	Property table	Graph and text	Map-like visualization & infobox
Faceted Navigation	Yes, in graph	Yes	No	Yes	No	No	No	No
Query Model	Lookup	Dimension element selection	Manual selection	Lookup	Lookup	Lookup	Lookup	Lookup
Query suggestion	Yes	No	No	No	No	No	No	No
LD editing utilities	Yes, Ontology verification	Yes, add/edit/delete properties to data, create/edit/import knowledge bases	No	Yes, add/edit/delete nodes/properties	No	Yes, Annotate using DBpedia Spotlight	No	No
LD View Save	No	No	No	Yes	No	No	No	No
Breadcrumb	No	Yes, sessions	Yes, sessions	Yes	No	No	No	No
Live Endpoints	Yes, Multiple	Yes	Yes	Yes	Yes	Yes	Yes	Yes
API	No	No	No	No	No	No	No	No

ESSs are more advanced systems compared to LDBs due to their advanced features. TABLE II presents a summary to ESSs studied in the previous section with additional features summarized from [12].

From the table, it can be noticed that the DBpedia dataset is used as a data resource by all but inWalk. In the case of Aemoo, SWOC, and LED, external sources are used to compute relations among dataset nodes. Search engines, microblogging systems, and social tagging methods are examples of external source. As for query model, the DBpedia lookup service provided by DBpedia SPARQL endpoint is utilized by the majority of ESSs.

The breadcrumb facility, also known as browsing history, is available in all systems through using browser memory sessions. This facility clearly decreases the server requests and supports users the ability to analyze and compare results with previous states. Result explanation feature support searchers to figure out how the results are presented, this utility is offered by Linked Jazz, Seevl, Aemoo and Discovery Hub in different visualizations such as graph and textual forms.

Results are usually ranked according to the similarity measures. inWalk, Linked Jazz, and Aemoo present their results as graphs so the rankings are not available, while the remaining systems present them in different ways. In SWOC, the node size of the graph shows the similarity to the queried concept, the bigger node, the most similar is. The top thumbnail sketch in a list is the most relevant result to the query; this can be seen in rankings of Seevl and Yovisto. LED uses the tag font size as the relevance to the query, and Discovery Hub categorizes the results based on the labels of resources in a list of thumbnail sketches.

TABLE II. OVERVIEW OF ESSS

System Name	Yovisto	SWOC	LED	Aemoo	Seevl	Discovery Hub	Linked Jazz	inWalk
Release Date	2009	2010	2010	2012	2012	2013	2013	2014
Web Address	www.yovisto.com	sisinflab.poliba.it/semantic-wonder-cloud	sisinflab.poliba.it/led	wit.istc.cnr.it/aemoo	play.seevl.fm	discoveryhub.co	linkedjazz.org	islab.di.unimi.it/inwalk
Main Data	DBpedia EN+DE	DBpedia EN	DBpedia EN	DBpedia EN	DBpedia, Freebase & MusicBrainz	DBpedia EN, FR, IT	DBpedia -> Linked Jazz DB	Freebase & Twitter
Auxiliary Data	No	Search Engines & Tagging Systems	Search Engines & Tagging Systems	External services	No	No	No	No
Query Model	Keyword search	DBpedia lookup	Keyword search	Lookup	Lookup	DBpedia Lookup	Manual selection from a list	Lookup & selection
Matching	String-match	Direct match (lookup)	String matching	Direct match (lookup)	Direct match (lookup)	Direct match (lookup)	Selection	Direct match (lookup)
Domain in (Purpose)	Academic Videos	IT Domain	ICT	General	Music	General	Jazz Musicians	Athletes or Twitter News
Database Method	Freebase Parallax	SWOC Storage	LED Storage	Knowledge Pattern (KP) Repository Manager	Scalable RDF/ NoSQL storage by OpenLink Virtuoso	Virtuoso, MySQL	Linked Jazz Name Directory	inWalk repository
Database Purpose	map queries to entities	Stores popularity & similarity values for pairs of resources	Stores results computed by Ranker	responsible for the storage, indexing & fetching of KPs	Scalability	User account & follows	Stores individuals represented by literal triples	Provides a high-level view of relevant LD
Principal Layout	Query suggestions	Graph	Tags cloud	Graph	List	List	Graph	Graph
Results Explanations	No	No	No	Wikipedia-based	Shared properties	Yes, Text & Graph	Yes, Interview Transcripts	No
Breadcrumb	Sessions & registration	Sessions	Sessions	Sessions	Sessions & registration	Sessions & registration	Sessions	Sessions
Algorithms	Set of heuristics	DBpedia Ranker	DBpedia Ranker	EKP filtered view	LDSD, DBrec algorithm	Semantic spreading activation	Mapping & Curator Tool & the Transcript Analyzer	HCf^{+} clustering algorithm
Ranking	Yes	Yes, graph size	Yes	No	Yes	Yes	No	No
Offline Processing	Yes	Yes, similarity of pairs	Yes	Yes , EKP part	Yes	No, on the fly	Yes, Linked Jazz DB	Yes
API	RDF triple-store	No	Yes, RESTful	Yes, RESTful	Yes, Content negotiation JSON-LD	No	Yes, JSON, RDF & GEXF files	No
Faceted Navigation	Yes	No	No	No	Yes	Yes	No	No
User Interface	HTML	Flash based	HTML	HTML	HTML, ajax	HTML	HTML5 + jQuery	HTML5 + JS

Different algorithms are used in the ESSs to calculate the similarity values. SWOC and LED use DBpediaRanker to calculate the similarities between DBpedia resources. A collection of heuristics to decide the best similar entity in the dataset is utilized in Yovisto. EKP has been used in Aemoo to determine the related type of resources and to state the representative classes to clarify entities of a particular type. Seevl employs LDSD and DBrec algorithms to bring musical recommendations. A semantic-sensitive traversal algorithm combined with a graph sampling technique is the base algorithm of Discovery Hub. In Linked Jazz, Mapping and Curator tools were used to construct its dataset and refine its data respectively. The inWalk employs HC^{I+} clustering algorithm for the construction of inCloud.

Faceted browsing feature, which allows users to filter the results, is proposed by Seevl, Yovisto and Discovery Hub.

Database technology has been used in most of the systems with different aims. Freebase Parallax is utilized in Yovisto so as to get a collection of entities first, and retrieve videos related to entities. Both systems SWOC and LED benefit from the database technology by using a Storage to keep the similarity values among pairs of nodes in a DBMS so the retrievals of the system are more efficient at runtime. Knowledge Pattern Repository Manager is utilized by Aemoo

to store, index and fetch retrievals from Knowledge Patterns. OpenLink Virtuoso platform for RDF/NoSQL data is used by Seevl for the aim of scalability.

Regarding the application program interface (API) of the systems, five of the systems, Yovisto, LED, Aemoo, Seevl and Linked Jazz, provide facilities for mashup Web applications. Yovisto published its metadata in RDF format, embedded in web pages as RDFa and reachable via an RDF triple store. LED tag cloud generation is publically available as RESTful web service. The server-side section of Aemoo is released based on a REST Web service in Java, while its client side interface interacts with third party sections by REST interfaces via AJAX. For the sake of building applications on top of Seevl, Content Negotiation[25] on Seevl server is enabled and provides all the data as JSON-LD. The outputs of the Linked Jazz API are JSON, RDF Triples, and Gephi GEXF files. Although the default return is JSON, some of the ESSs allow the data return in other formats.

Fig. 1 describes statistics of the features used by all of the systems reviewed in this paper. As it can be noticed, DBpedia has been used as the dataset by 88% of the systems, which indicates the importance of this dataset and its wide usage. DBpedia Lookup Service has been used as query model by 63% of the systems, however, all of the systems that use DBpedia as their main dataset, use DBpedia Lookup Service as their query model. The presentation of the faceted navigation, which helps the users to realize and comprehend the information space, is still poor, this is due to the heavy processing of large-sized datasets.

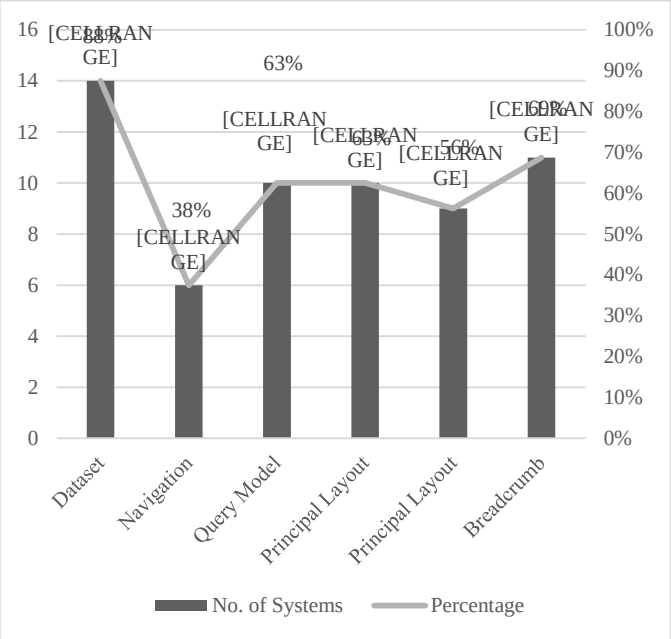


Fig. 1. Common features used by the systems

V. Conclusion

The effortless process of sharing data online created a massive volume of data which is increasing continuously. Consequently, the information retrieval systems are facing new challenges in order to return relevant data. The growth of the syntactic Web to the SW technology where the information is understandable by machines has raised expectations. This technology improves the efficiency of searching systems through automatic processing of data.

This paper has given an overview of the SW technology and LD. The concept of LDBs and ESSs are clarified. Afterward, the most recent LDBs and ESSs for LD have been reviewed in details. The motivation behind this study was to provide the reader with a clear explanation of the LDBs and search systems such as the way they use LD, their query paradigm, principal layout and used algorithms.

The future prospects of the exploration systems based on LD are promising. Moreover, the interest in this area will continue to expand constituting a critical enhancement for the future of the search experience and its results.

[25] http://en.wikipedia.org/wiki/Content_negotiation

References

[1] G. Madhu, D. A. Govardhan, and D. T. Rajinikanth, "Intelligent Semantic Web Search Engines: A Brief Survey," ArXiv Prepr. ArXiv11020831, 2011.

[2] G. Marchionini, "Exploratory Search: From Finding to Understanding," Commun ACM, vol. 49, no. 4, pp. 41–46, Apr. 2006.

[3] R. Mirizzi, A. Ragone, T. D. Noia, and E. D. Sciascio, "Semantic Wonder Cloud: Exploratory Search in DBpedia," in Current Trends in Web Engineering, F. Daniel and F. M. Facca, Eds. Springer Berlin Heidelberg, 2010, pp. 138–149.

[4] T. Jiang, "Exploratory Search: A Critical Analysis of the Theoretical Foundations, System Features, and Research Trends," in Library and Information Sciences, Springer, 2014, pp. 79–103.

[5] J. A. R and M. Kurian, "A Survey on Tools essential for Semantic web Research," Int. J. Comput. Appl., vol. 62, no. 9, pp. 26–29, Jan. 2013.

[6] T. Berners-Lee, J. Hendler, and O. Lassila, "The semantic web," Sci. Am., vol. 284, no. 5, pp. 28–37, 2001.

[7] D. Fensel, Spinning the Semantic Web: bringing the World Wide Web to its full potential. MIT Press, 2005.

[8] K. Krieger and D. Rosner, "Linked Data in E-Learning: A Survey," Semantic Web 0, pp. 1–9, 2011.

[9] T. Heath and C. Bizer, Linked Data: Evolving the Web into a Global Data Space, 1st edition. Morgan & Claypool., 2011.

[10] Le Hors, M. Nally, and S. Speicher, "Using read/write Linked Data for Application Integration-Towards a Linked Data Basic Profile.," presented at the LDOW, 2012.

[11] Bizer, T. Heath, and T. Berners-Lee, "Linked data-the story so far," Int. J. Semantic Web Inf. Syst., vol. 5, no. 3, pp. 1–22, 2009.

[12] N. Marie and F. Gandon, "Survey of linked data based exploration systems," presented at the IESD 2014-Intelligent Exploitation of Semantic Data, 2014.

[13] Geroimenko and C. Chen, Visualizing the semantic web: XML-based internet and information visualization. Springer Science & Business Media, 2006.

[14] D. V. Camarda, S. Mazzini, and A. Antonuccio, "Lodlive, exploring the web of data," presented at the Proceedings of the 8th International Conference on Semantic Systems, 2012, pp. 197–200.

[15] P. E. R. Salas, M. Martin, F. M. D. Mota, S. Auer, K. Breitman, and M. Casanova, "Publishing statistical data on the web," presented at the Semantic Computing (ICSC), 2012 IEEE Sixth International Conference on, 2012, pp. 285–292.

[16] J. M. Brunetti, S. Auer, R. García, J. Klímek, and M. Nečaský, "Formal linked data visualization model," presented at the Proceedings of International Conference on Information Integration and Web-based Applications & Services, 2013, p. 309.

[17] E. H. Chi, "A taxonomy of visualization techniques using the data state reference model," presented at the Information Visualization, 2000. InfoVis 2000. IEEE Symposium on, 2000, pp. 69–75.

[18] Micsik, Z. Tóth, and S. Turbucz, "LODmilla: Shared Visualization of Linked Open Data," presented at the Theory and Practice of Digital Libraries--TPDL 2013 Selected Workshops, 2014, pp. 89–100.

[19] Y. Qu et al., "SView: Smart Views for Browsing Linked Entities," Semantic Web Chall., 2014.

[20] D. Lukovnikov, C. Stadler, and J. Lehmann, "LD viewer-linked data presentation framework," presented at the Proceedings of the 10th International Conference on Semantic Systems, 2014, pp. 124–131.

[21] Vagliano, M. Marengo, and M. Morisio, "DBpedia Mobile Explorer," presented at the Research and Technologies for Society and Industry Leveraging a better tomorrow (RTSI), 2015 IEEE 1st International Forum on, 2015, pp. 181–185.

[22] D. Auber, C. Huet, A. Lambert, B. Renoust, A. Sallaberry, and A. Saulnier, "GosperMap: Using a gosper curve for laying out hierarchical data," Vis. Comput. Graph. IEEE Trans. On, vol. 19, no. 11, pp. 1820–1832, 2013.

[23] F. Valsecchi, M. Abrate, C. Bacciu, M. Tesconi, and A. Marchetti, "DBpedia Atlas: Mapping the Uncharted Lands of Linked Data," Proc. Workshop Linked Data Web, 2015.

[24] R. W. White and R. A. Roth, Exploratory Search: Beyond the Query-Response Paradigm Synthesis Lectures on Information Concepts, Retrieval, and Services. Morgan & Claypool Publishers, 2009.

[25] R. W. White, B. Kules, and B. Bederson, "Exploratory search interfaces: categorization, clustering and beyond: report on the XSI 2005 workshop at the Human-Computer Interaction Laboratory, University of Maryland," presented at the ACM SIGIR Forum, 2005, vol. 39, pp. 52–56.

[26] V. Dimitrova, L. Lau, D. Thakker, F. Yang-Turner, and D. Despotakis, "Exploring Exploratory Search: A User Study with Linked Semantic Data," in Proceedings of the 2Nd International Workshop on Intelligent Exploration of Semantic Data, New York, NY, USA, 2013, p. 2:1–2:8.

[27] K. Jacksi, N. Dimililer, and S. R. Zeebaree, "A SURVEY OF EXPLORATORY SEARCH SYSTEMS BASED ON LOD RESOURCES," Proc. 5th Int. Conf. Comput. Inform. ICOCI 2015, pp. 501–509, 2015.

[28] J. Waitelonis and H. Sack, "Towards Exploratory Video Search Using Linked Data," Multimed. Tools Appl, vol. 59, no. 2, pp. 645–672, Jul. 2012.

[29] R. Mirizzia, A. R. T. Di Noiaa, and E. Di Sciascioa, "Lookup, Explore, Discover: how DBpedia can improve your Web search," 2010.

[30] Musetti et al., "Aemoo: Exploratory search based on knowledge patterns over the semantic web," Semantic Web Chall., 2012.

[31] Passant, "seevl: mining music connections to bring context, search and discovery to the music you like," in Semantic Web Challenge, 2012.

[32] N. Marie, F. Gandon, M. Ribière, and F. Rodio, "Discovery Hub: On-the-fly Linked Data Exploratory Search," in Proceedings of the 9th International Conference on Semantic Systems, New York, NY, USA, 2013, pp. 17–24.

[33] M. C. Pattuelli, M. Miller, L. Lange, S. Fitzell, and C. Li-Madeo, "Crafting Linked Open Data for Cultural Heritage: Mapping and Curation Tools for the Linked Jazz Project," Code{4}lib, no. 21, Jul. 2013.

[34] S. Castano, A. Ferrara, and S. Montanelli, "inWalk: Interactive and Thematic Walks inside the Web of Data.," presented at the EDBT, 2014, pp. 628–631.

[35] A. Ferrara, L. Genta, and S. Montanelli, "Linked Data Classification: A Feature-based Approach," in Proceedings of the Joint EDBT/ICDT 2013 Workshops, New York, NY, USA, 2013, pp. 75–82.

Qos-based Computing Resources Partitioning between Virtual Machines in the Cloud Architecture

Evgeny Nikulchev
Moscow Technological Institute
Leninskiy pr., 38A, Moscow, Russia
119334

Evgeniy Pluzhnik
Moscow Technological Institute
Leninskiy pr., 38A, Moscow, Russia
119334

Oleg Lukyanchikov
Moscow Tecnological University
MIREA
Vernadsky Avenue, 78, Moscow,
Russia 119571

Dmitry Biryukov
Moscow Tecnological University MIREA
Vernadsky Avenue, 78, Moscow, Russia 119571

Elena Andrianova
Moscow Tecnological University MIREA
Vernadsky Avenue, 78, Moscow, Russia 119571

***Abstract*—Cloud services have been used very widely, but configuration of the parameters, including the efficient allocation of resources, is an important objective for the system architect. The article is devoted to solving the problem of choosing the architecture of computers based on simulation and developed program for monitoring computing resources. Techniques were developed aimed at providing the required quality of service and efficient use of resources. The article describes the monitoring program of computing resources and time efficiency of the target application functions. On the basis of this application the technique is shown and described in the experiment, designed to ensure the requirements for quality of service, by isolating one process from the others on different virtual machines inside the hypervisor.**

Keywords—cloud computing architecture; simulation; software for monitoring computer resources

I. INTRODUCTION

Cloud services are one of the most popular emerging areas of modern IT industry. The development of cloud computing technology has set a number of specialized tasks requiring fundamental results. Specifically, the load control task to provide the required quality of service [1]. The allocation of resources between virtual machines and remote data centers with unknown parameters of data channels requires dynamic control of the operational parameters of virtualization and data transfer quality. Introduction of feedback from controlled parameters allows making corrective actions; on the other hand, a task of optimal control can occupy a substantial part of computing processes and communication channels [2, 3]. It is possible to get feedback from the hypervisors, on which modern cloud systems are built, and from the cloud-based software itself, if developers provided such opportunities. The idea is proposed to develop theoretical basis for the program resource management in distributed cloud infrastructures, including the methods, models, patterns and the prototype of software middleware.

Managing information systems based on full use of cloud infrastructure offers the solution for task of creating a platform that automates the allocation of resources at the lowest cost [4, 5].

Main directions:

- guaranteeing the quality of service (Quality of Service, QoS);
- optimizing resources (reduction of energy consumption, cost optimization, etc.);
- providing security (to guarantee confidentiality and data integrity).

In general, global trends are such that cloud services are replacing classical architecture of information systems. Therefore we should be prepared for the transfer of existing systems to the cloud [6].

In the cloud technology, a duplication of program code is used to ensure reliability of data transmission. In case of container technologies, only code libraries are duplicated. Some container technology allows avoiding such duplication of system libraries code. The main negative effect of code duplication effect is the heavy load on the CPU cache. The report [7] carries out a detailed analysis of this effect. In practice, the increase of the processor cache use leads to a significant drop of system performance.

One of the features of virtual machines - the lack of direct access to physical memory of the main system. This is one of the reasons that hinder interoperability of the system with input-output devices [8].

For large data storage it is common to split data into smaller segments, which reduces the average access time. This acceleration is achieved by reducing the required number of read operations, during each of them a continuous segment of data can be read from the disk. In all these systems, data storage uses a local file system or a relational database, which significantly limits the possibilities of scaling. When using cloud computing alternative for read operation is establishing a connection and accessing the object in cloud storage. In cloud systems use semi-structured data, hierarchical models, noSQL

system and others. Many researchers made it advisable to use graph data representation models [9, 10]. An important feature of the graph methods is the existence of a significant number of polynomial solutions.

For modern cloud applications data transmission networks is a bottleneck. And one of the most important tasks in ensuring the functioning of the systems - configuration of the load in networks, to provide the use of applications in the cloud.

An algorithm for constructing a module of software configuration developed based on dynamic models [11, 12] The review [13] describes the basic QoS assurance system at the network level.

The cloud system, under the conditions of availability of computing resources on the server, there is a static load balancing, in which the distribution is carried out in advance. In the conditions of computing resources constraints and an ever-changing number of requests in the system, static load balancing does not give effect for heavy loaded systems, requiring dynamic load balancing.

Hypervisors emulate almost all the equipment, creating their virtual copies. Parameters of the virtual processor copies, memory, disk, and network adapter, affect the computing performance. Managing this equipment, you can change the virtual machine performance.

Virtual resources are considered to be infinite, but in practice there are physically limited resources on which hypervisor runs. Therefore, the optimal distribution of the physical computing resources between the virtual machines is an important task.

For each virtual machine, the required amount of computing resources is allocated – RAM, the number and frequency of processors, etc. Those are so-called configuration settings of virtual machine. This virtual machines are not always hold all the power selected for their computing, sometimes they are idle, when there are no requests, so at this time, these resources can be used by other virtual machines. This configuration in the hypervisor VMWare ESXi includes a number of parameters, such as Limit, Reservation and Shares for the Resource Pool within the VMware DRS cluster and ESX hosts. It is these three parameters that determine the memory consumption by the virtual machines and CPU resources of VMware ESX host allocated to them.

II. Formulation of the Problem

Limit determines the limit of consumption of physical resources by the virtual machine pool. Thus, if this parameter is set within the physical resources, there are no conflicts occurred between the virtual machines, but if this parameter is the same for some of the virtual machines, then conflicts may arise in heavy loaded systems. If both processors require more resources than was allocated by the physical parameter Limit, they just start to wait when one processor is released, thus time delays occure.

If memory usage reaches the parameter Limit, then it goes to the *swap* area, which, of course, slower, also causing delays.

The *Shares* parameter defines the prioritization of consumption by the virtual machines among each other within the ESX host and the resource pool. The three standard settings High, Medium and Low priority indicate the priority ratio.

Reservation parameter for a virtual machine or resource pool determines how much physical memory or CPU resources will be guaranteed to the virtual machine during operation. If the virtual machine has not yet reached Reservation parameter, the unused resources may be granted to other virtual machines (in Shares); otherwise, if it has already reached that level, then its computing resources will no longer be re-allocated.

Optimally configuring and managing these parameters of computational resources allocation between virtual machines, you can optimize the use of the physical resources of the server by reducing their downtime.

Dynamic load balancing is mainly divided into several sub-tasks: initial allocation of resources; evaluation download compute nodes; initiation of load balancing; taking decisions about balancing; moving objects (migration).

Thus, the problem lies in the fact that the resources of cloud computing environment can perform all the current load with minimal loss of performance. To dynamically manage cloud infrastructure it was suggested to use feedback. For the cloud technology researchers suggest different ways of constructing a system of dynamic equations. Providing a feedback is possible by hypervisors on which modern cloud systems are built, and application software, if such opportunities were provided by developers.

The hypervisor acts as an object that generates a signal that contains the parameters for monitoring the workload of processors, network server's memory and each virtual machine individually. The control mechanism is a software hypervisor management tool, which is based on one or more specified actions that determine the law (algorithm) of management. It generates a control action for the server and maintains a predetermined level or changes the state according to a certain law, which can be displayed on the corresponding output signal.

In almost all applications logging functionality is always provided, through which you can get the time taken by operations. If cross-platform standardized service syslog is used for logging, then with its help you can also send information to the application to monitor and manage the hypervisor (Fig. 1).

As a result, it is possible to display all load parameters for computing resources and query processing time, which allows detecting complex queries to redirect them to the available virtual machines.

A number of studies on the establishment of management systems based on non-linear models with control was conducted. As a controlled parameter process performance ratio is used depending on specified priorities. These parameters have been proposed as promising approaches for achieving QoS with unpredictable load conditions. The purpose was to control the allocation of resources so that the current balance of performance is in line with the priority given

to relations under the given constraints. Also, a similar system is implemented to control traffic.

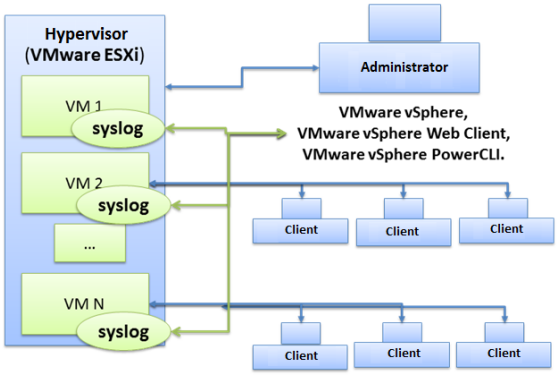


Fig. 1. The infrastructure for monitoring and managing the hypervisor

It should be borne in mind that the introduction of feedback allows to make corrective actions, but at the same time occupies part of computing resources and communication channels, so that is necessary for optimal control problems, where the quality criteria can minimize the processing time and resource constraints. Based on studies on the experimental stand, situations often occur, in which used control method can provide a guaranteed quality of service. These situations include the following: the addition of a new service in the software, which has become very popular, high growth in data volume, need to increase the number of virtual machines, the growing number of geographically dispersed users who require additional cache servers.

III. Simulation

The cloud system, under the conditions of availability of computing resources on the server, there is a static load balancing, in which the distribution is carried out in advance. For this distribution the experience of previous systems and test results are often taken into account. But in the conditions of computing resources constraints and an ever-changing number of requests in the system, static load balancing does not give effect for heavy loaded systems, requiring dynamic load balancing. Dynamic load balancing is essentially divided into several sub-tasks [12, 14, 15]:

- The initial allocation of resources;
- Evaluation of compute nodes load;
- Initiation of load balancing;
- Decision-making about balancing;
- Moving objects (migration).

The problem is in providing enough resources of cloud computing environment ($S_{обл}$) to perform all the current system load with minimal loss of productivity [5]. To solve this problem, monitoring of the system is used, based on which the management is organized.

The formal criterion can be written as:

$$S_{cl} = F(L_{mem}, L_D, L_{cp}, L_{nw}, T_D) \rightarrow max \quad (1)$$

Here L_{mem} is the memory resource, L_D is the drive resource, L_{cp} is the CPU resource, L_{nw} is the network resource, T_D response from drive system.

To assess the resources at cold start of virtual machines in the cloud system, streams are generated to simulate the users' requests with a given intensity.

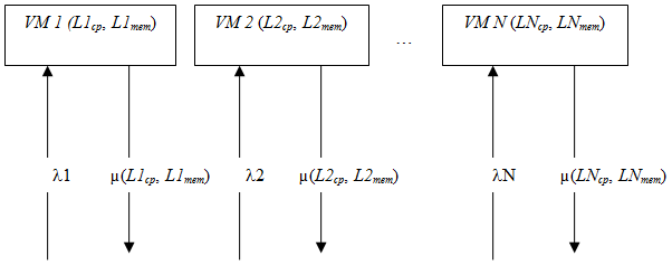


Fig. 2. A simulation model for the processing of cloud system applications

Fig. 2 shows a server with CPU frequency L_{cp} and RAM size of L_{mem}, i.e. Lцп and Loп are limitations of a physical server. Hypervisor is installed on the server with deployed virtual machines, that perform various tasks of processing user requests with incoming intensities λ1, λ2, .., λN. Time of processing the user requests depends on the resources, allocated to virtual machines VM 1 ($L1_{cp}$, $L1_{mem}$), VM 2 ($L2_{cp}$, $L2_{mem}$), ... VM N (LN_{cp}, LN_{mem}). It also provides the desired level of service. It turns out, the intensities of query processing for users μ(L1цп, L1оп), μ(L2цп, L2оп), ..., μ(LNцп, LNоп) are functions, dependent on allocated for virtual machine CPU and memory resources. It is possible to get empirically the function of the intensity of processing user requests through a series of experiments, timing the execution time of queries with different configurations, and then calculating the μ by the formula:

$$\mu = \frac{1}{\bar{t}}. \quad (2)$$

where $\bar{t}$ is the average time of processing a user query.

The main purpose of the cloud system is to ensure the required quality of service, and thus to ensure the minimum time to process the user queries (t_q):

$$\sum_{i=1}^{N} \bar{t}_{iq} \rightarrow min. \quad (2)$$

Since $\bar{t_q} = \frac{\bar{r}}{\lambda}$, where $\bar{r}$ is the average number or queries in a queue, then it should be minimal as well:

$$\sum_{i=1}^{N} \bar{r}_i \rightarrow min \quad (3)$$

If we consider each virtual machine as a single-channel Queueing System (QS) with endless queues, the average number of requests in the queue is calculated as follows: [11]

$$\bar{r} = \frac{\rho^2}{1-\rho}, \text{ where } \rho = \frac{\lambda}{\mu(LN_{cp}, LN_{mem})} < 1. \quad (4)$$

The model of QS depends on problems solved in each virtual machine. If requests can be processed in parallel on a virtual machine, the system should be considered as a multi-channel QS with infinite queue, where the number of channels is determined by the amount of processors allocated to the virtual machine. If processing the query involves all the

kernels, the system is a single-channel QS, but the function of processing applications will get another dependency, which is the number of processors.

In the QS with infinite queue, very important condition (4) would be the intensity of the receipt of applications, that has to be less than the intensity of the processing of applications. Otherwise, queue will grow indefinitely. Therefore, when choosing LNцп and LNоп parameters, it determines the minimum search threshold of optimal solutions for (2) and (3), therefore solution obtained in (4) represents the parameters of the Reservation.

When computing resources of the server are enough to perform the tasks

$$\sum_{i=1}^{N} L_{icp} < L_{cp} \quad и \quad \sum_{i=1}^{N} L_{imem} < L_{mem}\,, \qquad (5)$$

remaining resources can be identified in the Share for all virtual machines, which will automatically perform the balancing of the system by a hypervisor, increasing the average time of processing user requests.

$$Share_{cp} = L_{cp} - \sum_{i=1}^{N} L_{icp}\,,\ Share_{озу} = L_{оп} - \sum_{i=1}^{N} L_{iоп}. \qquad (6)$$

When computing resources are not enough to perform the tasks and provide the required quality of service to users

$$\sum_{i=1}^{N} L_{icp} > L_{cp} \quad and \quad \sum_{i=1}^{N} L_{imem} > L_{mem}\,, \qquad (7)$$

it is necessary to apply the methods of dynamic cloud management system.

IV. Experiments and Results

Experiments were carried out with the help of the developed software, aimed to control computing resources and run-time operations on a computing server with multiple virtual machines, on which loading applications are installed.

The software on the administrator's workstation performs centralized collection of characteristics of virtual machines, the hypervisor, in addition to time required for processing the system operations. Displaying all of this information on a chronological schedule will allow the administrator to decide on a cloud infrastructure management, as shown in Fig. 3 (red - the CPU usage for VM1; lilac – RAM usage for VM1; blue - CPU usage for VM2; green – RAM usage for VM2; crimson - runtime of the function 1; yellow – runtime of the function 2).

To show all the parameters on one chart it is necessary to normalize them. For the convenience, the parameters are divided into 2 types:

a) Computing parameters of virtual machines on the right y-axis, which are displayed as a percentage of the maximum values of the virtual machines.

b) Parameters of time required to perform various functions on the left y-axis. Each point of time run-time parameter is marked at the time of the completion of the function, but knowing its time, you can calculate its operating range. Orange in Fig. 3 is a range of work completed in a single function finished at 3:44:10 that runs 25 seconds, respectively, the implementation of which began in 3:43:45. This graph allows the operator to identify more complex queries or transactions, and redirect or move applications to available virtual machines.

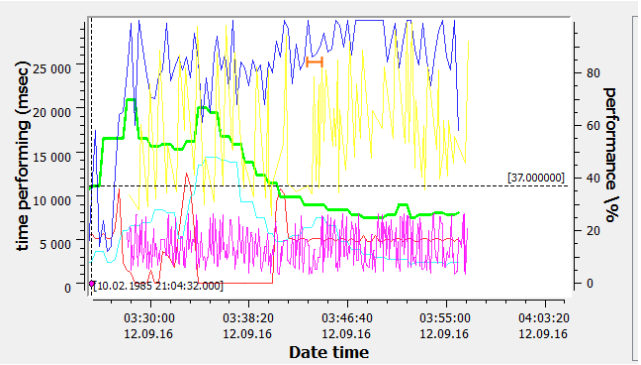


Fig. 3. Graphical display of execution time of functions and computational parameters of virtual machines

The developed method allows determining the display load unbalance moment and deciding on the load balancer. To demonstrate this, the following experiment was conducted.

The application was developed that emulates the load. It performs a recursive search for a file on disk. This application creates separate threads to search at random intervals within a predetermined range, and then closes them at random intervals. The search is creates a heavy load on the CPU computing resources, so other computing parameters will be ignored.

For the experiments a server HP ProLian ML 110 G6 with Intel® processor Xeon® CPU X3450 @ 2.67 Ghz and 4GB RAM was used, with hypervisor VMware ESXi and installed virtual machines running Ubuntu.

The experimental results are shown in Fig. 4, where the blue - CPU usage for VM1, lilac - RAM usage for VM1, red - RAM usage for VM2, green - CPU usage for VM2, raspberry - run-time function f1, which is run from 1 to 10 seconds, and carried out from 1 to 2 seconds, yellow - run time function f2, which is run from 10 to 60 seconds, and was carried out from 5 to 15 seconds, gray - f3 runtime function that starts from 30 to 120 seconds, and was performed for 60 seconds.

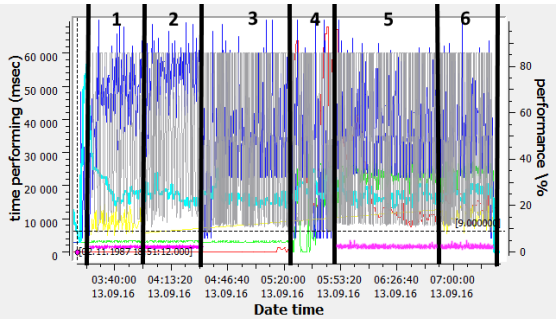


Fig. 4. The experimental results

The whole experience can be divided into six stages.

Two cores and 2GB of RAM are allocated to Virtual machine 1, all three functions are running on it. More results of the phase 1 are shown in Fig. 5. The average CPU usage was in the range of 60-90 percent, which could adversely affect the required quality of service to users. Hence there is a problem,

especially for labor-intensive operations, to identify and migrate them to other virtual machines in order to reduce CPU usage.

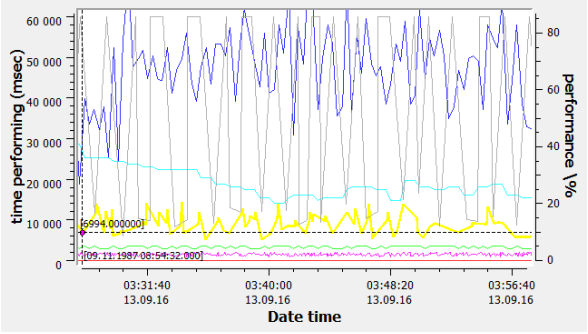


Fig. 5. Phase 1 experimental results

In the second phase, the results of which are shown in Fig. 6, function f2 has been switched off, which, as seen, has no effect on processor usage.

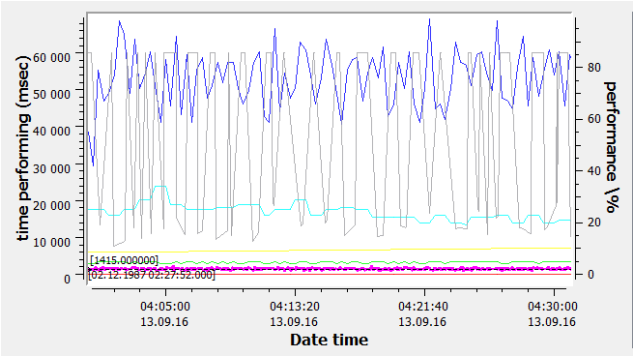


Fig. 6. Phase 2 experimental results

In the third phase, the results of which are shown in Fig. 7, function f1 has been switched of. It did reduce the usage of the processor in the range of 40-80 percent, this leads to the conclusion that the function f3 is the most labor-intensive.

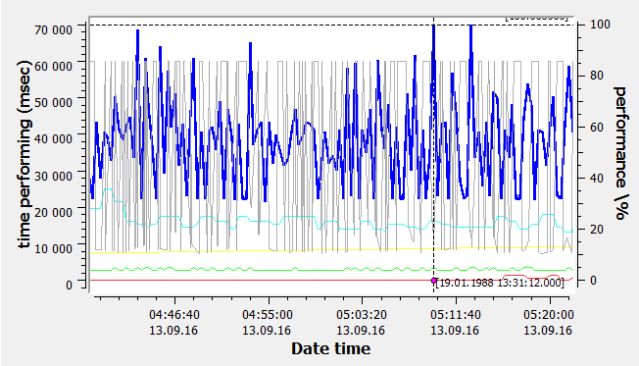


Fig. 7. Phase 3 experimental results

In phase 4 resources of virtual machine 1 were reallocated to another virtual machine, resulting in two identical virtual machines with a single processor core and 1GB of RAM.

In the fifth phase, the results of which are shown in Fig. 8, f3 function remained to work in virtual machine 1, and f1 function, was launched on virtual machine 2. And compared to phase 2 the gain in productivity was obtained in the first virtual machine.

In the last sixth phase, shown in Figure 9, the function f2 has been added to virtual machine number 2, which does not significantly impact the CPU usage on the second virtual machine.

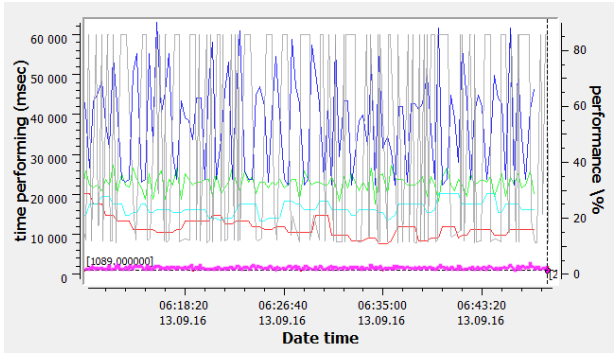


Fig. 8. Phase 5 experimental results

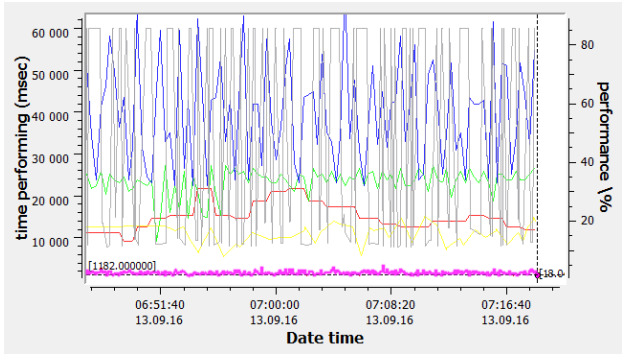


Fig. 9. Phase 6 experimental results

As a result, using the same amount of computing resources, insignificant gains were obtained in the performance of a virtual machine 1 running the function f3, while other functions f1 and f2 were isolated, thus providing the desired quality of service for at least these functions . If in Phase 1 f3 affected the processing of functions f1 and f2, then by isolating them makes it impossible, which has the positive effect on the required quality of service for users.

V. Discussion

Too frequent load balancing can lead to the case, where simulation model slows down. The costs of balancing itself may surpass the possible benefit from its implementation. Therefore, for productive balancing it is necessary to determine the time of its initialization.

It requires:

- Determine the time of the load imbalance.
- Determine the degree of needed balancing by comparing the potential benefits of its implementation and the cost of it.

Load imbalance can be determined synchronously and asynchronously.

In synchronous determination of imbalance, all processors (network computers) are interrupted at certain times of synchronization, and imbalance is determined by comparing the load on a separate processor with a total average load. In asynchronous determination of imbalance, each processor keeps a history of its usage. In this case, the moment of synchronization for determination of imbalance is absent. The background process that runs in parallel with the application calculates the imbalance.

Most of the dynamic load-balancing strategies can be classified as centralized or fully distributed. With centralized strategy, one computer collects global information on the status of the entire computer system and makes a decision about moving tasks between the computers.With fully distributed strategy, each processor performs load-balancing algorithm to exchange information on the status with other processors. Migration occurs only between neighboring processors.

VI. Conclusion

The article proposed a method of initial allocation of computing resources for the virtual machines in the hypervisor using a simulation model of the processing user requests, which solved the problems (2) and (3) under the conditions (4-6). The article also lists the software for collection and analysis of computational load of the virtual machines in the hypervisor and the time of the effectiveness of the objective functions of applications that handle user requests. Based on this data, the operator has the possibility to reallocate the computational resources of the hypervisor, thereby providing dynamic control for the system.

The experiment showed the effectiveness of the use of this application. The most labor-intensive function was found, which, in consequence, was isolated from the other functions, which ensured their required level of service. The above methods for configuring and managing hypervisors allow better use of computing resources of servers.

In the long term development of the use of these methods it is possible to develop a algorithm that performs automatic migration of services, or redirecting requests that will organize the distribution of computing in the misty system.

Acknowledgment

This work was partially funded by the basic part of state task of the Ministry of education and science of Russia for scientific project #792: Research and development of methods and algorithms for constructing open distributed information systems for various purposes. The work is funded by the Moscow Institute of Technology.

References

[1] A. Jarray, J. Salazar, A. Karmouch, J. Elias, and A. Mehaoua “QoS-based cloud resources partitioning aware networked edge datacenters,” In 2015 IFIP/IEEE International Symposium on Integrated Network Management (IM), 2015, pp. 313-320.

[2] R. D. C. Coutinho, L. M. Drummond, Y. Frota and D. de Oliveira, “Optimizing virtual machine allocation for parallel scientific workflows in federated clouds,” Future Generation Computer Systems, vol. 46, pp. 51-68, 2015.

[3] C. Papagianni, A. Leivadeas, S. Papavassiliou, V. Maglaris, C. Cervell o-Pastor and A. Monje, “On the optimal allocation of virtual resources in cloud computing networks”, IEEE Transactions on Computers, vol. 62, no. 6, pp. 1060-1071, 2013.

[4] J. Tordsson, R. S. Montero, R. Moreno-Vozmediano and I. M. Llorente “Cloud brokering mechanisms for optimized placement of virtual machines across multiple providers,” Future Generation Computer Systems, vol. 28, no. 2, pp. 358-367, 2012.

[5] D. F. Bari, R. Boutaba, R. Esteves, L .Z. Granville, M. Podlesny, M. G. Rabbani, Q. Zhang, and M. F. Zhani, “Data Center Network Virtualization: A Survey,” IEEE Communications Surveys and Tutorials vol. 15, no. 2, pp. 909-928, 2013.

[6] Z. Á. Mann, “Allocation of Virtual Machines in Cloud Data Centers—A Survey of Problem Models and Optimization Algorithms,” ACM Computing Surveys (CSUR), vol. 48, no. 1, 2015.

[7] V. Medina and J. M. García, “A survey of migration mechanisms of virtual machines,” ACM Computing Surveys (CSUR),vol. 46, no. 3, 2014

[8] J. Li, Q. Wang, D. Jayasinghe, J. Park, T. Zhu and C. Pu, “Performance overhead among three hypervisors: An experimental study using hadoop benchmarks,” In 2013 IEEE International Congress on Big Data, 2013, pp. 9-16.

[9] E. Nikulchev, E. Pluzhnik, et al. “Features Management and Middleware of Hybrid Cloud Infrastructures,” International Journal of Advanced Computer Science and Applications, vol. 7, no. 1, pp. 30-36. 2016.

[10] J. Tordsson, R. S., Montero, R. Moreno-Vozmediano, and I. M. Llorente “Cloud brokering mechanisms for optimized placement of virtual machines across multiple providers,” Future Generation Computer Systems, vol. 28, no. 2, pp. 358-367, 2012.

[11] A. Jarray, A. N. Quttoum, H. Otrok and Z. Dziong, “DDP: A dynamic dimensioning and partitioning model of virtual private networks resources” Computer Communications, vol. 35, no. 8, pp. 906-915, 2012.

[12] E. Pluzhnik, E. Nikulchev and S. Payain, “Optimal control of applications for hybrid cloud services,” 2014 IEEE World Congress on Services, 2014, pp. 458-461.

[13] A. A. Abbasi and M. Hussain, “A QoS Enhancement Framework for Ubiquitous Network Environments,” International Journal of Advanced Science and Technology, vol. 43, pp. 37–48, 2012.

[14] Y. Diao, N. Gandhi et al. “Using MIMO feedback control to enforce policies for interrelated metrics with application to the Apache web server,” IEEE 2002 Network Operations and Management Symposium, 2002, pp. 219 - 234.

[15] T. Patikirikorala, L. Wang, A. Colman, J. Han, “Differentiated performance management in virtualized environments using nonlinear control,” IEEE Transactions on Network and Service Management, vol. 12, no. 1, pp. 101-113, 2015.

Multiobjective Optimization for the Forecasting Models on the Base of the Strictly Binary Trees

Nadezhda Astakhova
Ryazan State Radio Engineering University
Ryazan, Russia

Liliya Demidova
Moscow Technological Institute
Ryazan State Radio Engineering University
Moscow, Russia

Evgeny Nikulchev
Moscow Technological Institute
Moscow,
Russia

***Abstract*—The optimization problem dealing with the development of the forecasting models on the base of strictly binary trees has been considered. The aim of paper is the comparative analysis of two optimization variants which are applied for the development of the forecasting models. Herewith the first optimization variant assumes the application of one quality indicator of the forecasting model named as the affinity indicator and the second variant realizes the application of two quality indicators of the forecasting model named as the affinity indicator and the tendencies discrepancy indicator. In both optimization variants the search of the best forecasting models is carried out by means of application of the modified clonal selection algorithm. To obtain the high variety of population of the forecasting models it is offered to consider values of the crowding-distance at the realization of the second optimization variant. The results of experimental studies confirming the use efficiency of the modified clonal selection algorithm on the base of the second optimization variant are given.**

Keywords*—forecasting model; strictly binary tree; modified clonal selection algorithm; multiobjective optimization; affinity indicator; tendencies discrepancy indicator

I. Introduction

The main problem dealing with the development of the forecasting models is the problem of the right choice of the best forecasting model. The forecasting model based on the strict binary trees (SBT) and the modified clonal selection algorithm (MCSA) [1, 2] is presented in the form of antibody, which is coded by a line of symbols randomly selected from the corresponding alphabets. This antibody can be transformed to the analytical dependence, which is used for forecasting of a time series (TS). Obviously, the correct selection of antibodies is very important for the effective use of the MCSA [1 – 6].

The traditional approach in the short-term forecasting models choice consists in the quality estimation of the forecasting models by means of the average forecasting error rate (*AFER*), calculated for the training data sequence. Herewith the *AFER* should be minimized [1 – 6]. However, the use of the *AFER* as the unique quality indicator of the forecasting model is not always sufficient to determine the best forecasting model. Often it is required to consider the additional quality indicators of the forecasting model, such as the compliance to the seasonal tendencies of TS, the compliance to the trend of TS, lack of emissions, complexity of the forecasting model, etc. [6]. It is expedient to use the additional quality indicator, which will allow estimating the general tendency of values' change of the known elements of TS (for example, the tendencies discrepancy indicator) along with the *AFER* [6]. It is possible to increase the efficiency of the forecasting models on the base of the SBT, using the multiobjective MCSA at the solution of the problem of the medium-term forecasting. Herewith the affinity indicator based on the AFER and the tendencies discrepancy indicator can be used in the role of the objective functions.

The rest of this paper is structured as follows. Section 2 presents the main ideas of the original MCSA. Section 3 details the multiobjective optimization variant for the MCSA. Experimental results comparing two optimization variants (with the original MCSA and with the multiobjective MCSA) follow in Section 4. Finally, conclusions are drawn in Section 5.

II. The Main Ideas of the Modified Clonal Selection Algorithm

The MCSA simulates the natural laws of the immune system functioning and provides the formation of quite complex analytical functions [1], [2]. The principles of developing forecasting models of *k*-order with the use of the MCSA were investigated in [2]. The MCSA allows forming an analytical dependence on the base of the SBT at an acceptable time expenses, that describes certain TS values and provides a minimum value of the affinity indicator *Aff* based on the *AFER*:

$$AFER = (100\%/(n-k)) \cdot \sum_{j=k+1}^{n} |(f^j - d^j)/d^j| \qquad (1)$$

where d^j and f^j are respectively the actual (fact) and forecasted values for the *j*-th element of the TS; n is the number of TS elements.

The possible variants for analytical dependences are presented in the form of antibodies *Ab* , which recognize antigens *Ag* (the TS values). The antibody *Ab* is selected as "the best one". It provides the minimum value of the affinity indicator *Aff* . The antibody coding is carried out by recording signs in a line. These signs are selected from three alphabets: the alphabet of arithmetic operations (addition, subtraction, multiplication and division) $Operation = \{'+','-','\cdot','/'\}$; the functional alphabet $Functional = \{'S','C','Q','L','E','\_'\}$,

where letters $'S','C','Q','L','E'$ define mathematical functions "sine", "cosine", "square root", "natural logarithm", "exhibitor", and the sign $'-'$ means the absence of any mathematical function, the alphabet of terminals $Terminal = \{'a','b',...,'z','?'\}$, where letters $'a','b',...,'z'$ define the arguments required analytical dependence and the sign $'?'$ defines a constant. The use of these alphabets provides a correct conversion of randomly generated antibodies into the analytical dependence. The structure of these antibodies can be described with the help of the SBT. The number of signs in the alphabet of terminals *Terminal* in the antibody *Ab* determines the maximal possible order K of the models with $K \geq k$, where k is the real model order, i.e. having the value of the element d^j in the forecasting TS at the *j*-th moment of time, K values of the TS elements can be used as: d^{j-K},…, d^{j-2}, d^{j-1} [1], [2].

The use of the SBT type, illustrated in Fig. 1, allows building the complex analytical dependence and provides high accuracy of the forecasting TS [1], [2].

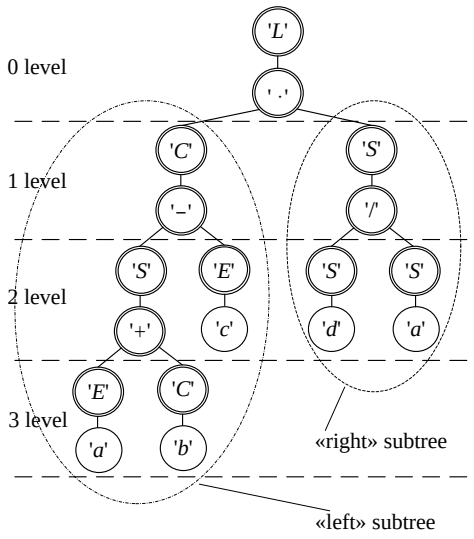


Fig. 1. An example of a strict binary tree, which is used to form antibodies

This SBT can be created as the composition of one "left" subtree of the maximum possible order $K = 3$ and some "right" subtrees of the maximum possible order $K = 2$. The term "left" ("right") subtree is used for the branch (left or right) of the SBT level where a new subtree should be included. It is rational to form antibodies by subdividing SBT into subtrees, then execute the subtree-walk of each vertex forming the ordered symbol lists on its vertices and then combining these lists consistently. Forming the symbol ordered list on the base of a subtree the consecutive double subtree-walk is carried out: at first moving the subtree bottom-up left to right it is necessary to bypass the vertices containing the alphabetic terminal signs *Terminal* in pairs and correspondingly above placed vertices containing the alphabetic functional symbols *Functional* and then moving in the same direction it is necessary to go around in pairs the vertices containing the alphabetic arithmetic operation signs *Operation* and correspondingly above placed vertices containing the alphabetic functional signs *Functional*. The first two signs of such antibody contain the pair of zero level SBT from the functional alphabet *Functional* and arithmetic operation alphabet *Operation*. Then there are the lists of the signs corresponding to the "right" maximum possible ordered subtrees $K = 2$ (moving the SBT bottom-up) and finally the symbol list of the «left» maximum possible ordered subtree $K = 3$.

For example, the antibody formed on the base of the SBT as shown in Figure 1 is coded by the line of signs: $L \cdot S/SeSdC - S + EaCbEa$, which can be transformed into the analytical dependence for the forecasting model with $k = 4$:

$$f(d^{j-1}, d^{j-2}, d^{j-3}, d^{j-4}) = ln(cos(sin(exp(d^{j-1}) + \\ + cos(d^{j-2})) - exp(d^{j-3})) \cdot sin(sin(d^{j-4}) / sin(d^{j-1})).$$

Interpreting the antibodies into the analytical dependences it is rational to use the recursive procedure of interpretation [2]. The MCSA applied to the searching for "the best" antibody defining "the best" analytic dependence includes the preparatory part (realizes the formation of the initial antibody population) and iterative part (presupposes the ascending antibodies ordering of affinity *Aff* the selection and cloning the part of "the best" antibodies, that are characterized by the least affine value *Aff* the hypermutation of the antibodies clones; self-destruction of the antibodies clones "similar" to the other clones and antibodies of the current population; calculating the affinity of the antibodies clones and forming the new antibodies population; suppression of the population received; generation of the new antibodies and adding them to the current population until the ingoing size; the conditional test of the MCSA completion).

III. Multiobjective Optimization

The average forecasting error rate *AFER*, which is also called the affinity indicator *Aff* (in the context of working with the MSCA) can be used as the first quality indicator for the forecasting models. The rate of discrepancy between the tendencies of two time series (the tendencies discrepancy indicator *Tendency*) can be used as the second quality indicator for the forecasting models [6].

The tendencies discrepancy indicator *Tendency* can be calculated as:

$$Tendency = h/(n - r - 1), \quad (2)$$

where h is the number of negative multiplications $(f^{j-1} - f^j) \cdot (d^{j-1} - d^j)$; $j = \overline{r+2, n}$; d^j and f^j are respectively the actual (fact) and forecasted values for the j-th element of TS; n is the number of TS elements; r is the model order; $n - r - 1$ is the total number of multiplications $(f^{j-1} - f^j) \cdot (d^{j-1} - d^j)$.

This indicator allows adapting the forecasting models on the base of the SBT and MCSA for the medium-term forecasting.

The affinity indicator (1) and the tendencies discrepancy indicator (2) must be used simultaneously at the quality assessment of the forecasting models on the base of the SBT and MCSA to solve the problem of the medium-term forecasting.

Various well proved approaches can be applied to the solution of the problem of the simultaneous accounting of two quality indicators for the development of the forecasting models. Herewith it is necessary especially to allocate approach, based on the several multiobjective optimization algorithms, including, evolutionary algorithms. In recent years a number of multiobjective evolutionary algorithms (MOEA) have been suggested [7] – [15]. The main reason for this is their ability to find the multiple Pareto-optimal solutions in one single simulation run. These algorithms work with a population of solutions. Therefore, the primary attention has to be paid to maintaining the diversity and spread of solutions. Such MOEAs provide a solution of the account's problem of the several objective functions (quality indicators) at the analisys of various applied problems. The multiobjective genetic algorithms (MOGA) [7] – [11] are the most known algorithms of the multiobjective optimization. It is necessary to say about the multiobjective clonal selection algorithms (MOCSA) [12] – [15]. However, these algorithms are less designed and, in the majority, borrow the principles of multiobjective optimization underlain in the genetic algorithms. The possibility of this loan can be explained with many similar mechanisms of the evolutionary process realization in the MOGA and MOCSA. The analisys of merits and demerits of the MOEAs shows that such the MOGAs as the NSGA-II and the NSGA-III are significantly better than othes because they can successfully solve more difficult problems of the multiobiective optimization [6].

In this regard the decision on expediency of the adaptation of the ideas put in the NSGA-II at the realization of the multiobjective MCSA which is applied for the selection of the forecasting models on the base of the SBT had been made. Herewith, it is necessary to understand the forecasting model (and the antibody corresponding to it) as the decision, and the quality indicator of the forecasting model as the objective function at the realization of the multiobjective optimization algorithm. All forecasting models with use of the notion "Pareto-dominance" can be divided to dominated and nondominated models [6].

Let $Q_{s,v}$ be a value of the v-th quality indicator for the s-th forecasting model ($v=\overline{1,V}$; $s=\overline{1,S}$); let V be a quantity of the quality indicators of the forecasting model; let S be a quantity of the forecasting models. The s-th forecasting model is dominated by the z-th forecasting model ($s=\overline{1,S}$; $z=\overline{1,S}$), if the following conditions are satisfied: the s-th forecasting model is dominated by the z-th forecasting model, if the following conditions are satisfied for all quality indicators: $Q_{s,v} \geq Q_{z,v}$ ($v=\overline{1,V}$), and also there is at least one the v^*-th ($1 \leq v^* \leq V$) indicator for which the condition $Q_{s,v^*} > Q_{z,v^*}$ is satisfied. A herewith all quality indicators must be minimized. The rank R_s must be calculated for every s-th forecasting model ($s=\overline{1,S}$). The rank R_s is equal to the quantity of the forecasting models which dominate over the s-th forecasting model. The rank R_s of the s-th nondominated forecasting model is equal to zero [6]. Let $V=2$, $Q_{s,1}=Aff_s$, $Q_{s,2}=Tendency_s$, where Aff_s and $Tendency_s$ are the values of the affinity indicator (1) and the tendencies discrepancy indicator (2) for the s-th forecasting model ($s=\overline{1,S}$) accordingly. The s-th forecasting model is dominated by the z-th forecasting model ($s=\overline{1,S}$; $z=\overline{1,S}$), if the following conditions are satisfied: ($Q_{s,1} \geq Q_{z,1}$ and $Q_{s,2} > Q_{z,2}$) or ($Q_{s,1} > Q_{z,1}$ and $Q_{s,2} \geq Q_{z,2}$), that is ($Aff_s \geq Aff_z$ and $Tendency_s > Tendency_z$) or ($Aff_s > Aff_z$ and $Tendency_s \geq Tendency_z$) [6].

The crowing distances τ_s ($s=\overline{1,S}$) can be calculated using the following algorithm [10, 11].

Step 1. To calculate ranks for all forecasting models in the population. To unite the models with identical values of the rank into one group.

Step 2. For every group of the forecasting models:

- to sort the forecasting models according to each quality indicator value in ascending order of magnitude;
- to assign the infinite distance to boundary values of the forecasting models in the group, i.e. $\tau_1=\infty$ and $\tau_{G_w}=\infty$, where G_w is the quantity of the forecasting models in the w-th group ($w=\overline{1,W}$); W is the groups' quantity; to assign $\tau_s=0$ for $s=\overline{2,G_w-1}$;
- to calculate the the crowding distance τ_s as:

$$\tau_s=\sum_{v=1}^{V}(Q_{s-1,v}-Q_{s+1,v})/(Q_v^{max}-Q_v^{min}), \qquad (3)$$

where $Q_{s-1,v}$ and $Q_{s+1,v}$ are the values of the v-th quality indicator ($v=\overline{1,V}$) for the forecasting models with the numbers ($s-1$) and ($s+1$), which are the nearest "neighbors" for the s-th model; Q_v^{min} and Q_v^{max} are the minimum and maximum values of the v-th quality indicator ($v=\overline{1,V}$) accordingly.

Fig. 2 shows how we can calculate the crowding distance on the base of two quality indicators. The points, marked with solid circles, correspond to the models with the minimum (zero) value of the rank (i.e. these points correspond to the Pareto front with the zero rank). To calculate the crowing distance for the s-th forecasting model it is required to define values of both quality indicators for the ($s-1$)-th and the ($s+1$)-th models, which are the nearest "neighbors" for the s-th model and have the same rank. Also, it is necessary to define the best and worst values of each quality indicator.

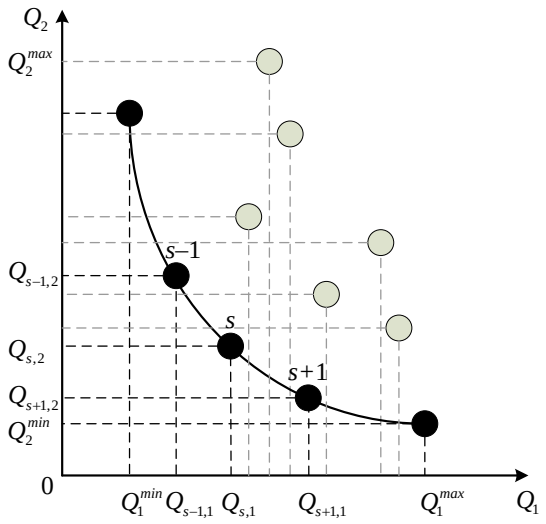


Fig. 2. The points used for the crowding distance calculation An example of a strict binary tree, which is used to form antibodies

The crowing distances τ_s ($s=\overline{1,S}$) for the s -th forecasting model on the base of two quality indicators can be calculated as [6]:

$$\tau_s = (Q_{s-1,1} - Q_{s+1,1})/(Q_1^{max} - Q_1^{min}) + \\ + (Q_{s-1,2} - Q_{s+1,2})/(Q_2^{max} - Q_2^{min}) \quad (4)$$

At the realization of the multiobjective MCSA the s -th forecasting model will be better than the z -th forecasting model, if:

$$R_s < R_z \text{ or } (R_s = R_z \text{ and } \tau_s > \tau_z).$$

If the s -th forecasting model is better than the z -th forecasting model, the s -th forecasting model is the candidate for transfer into the new generation.

For confirmation of prospects of the offered transformation of the MCSA it is offered to realize the following multiobjective optimization algorithm [6].

Step 1. To generate initial population of antibodies. Each antibody is coded on the base of the SBT and represents some forecasting model.

Step 2. To perform the nondominated sorting to population of antibodies on the base of two indicators of quality for the forecasting model (the affinity indicator (1) and the tendencies discrepancy indicator (2)).

Step 3. To choose the parents-antibodies for the next generation of the clones-antibodies based on the values of the rank and crowding distance.

Step 4. To pass to step 5 if desirable values of the quality indicators are reached or the quantity of generations in the MCSA is settled. Otherwise to pass to step 2.

Step 5. To accept the antibody with the minimum value of the affinity indicator (1) in the last population as the optimum decision. To use the forecasting model corresponding to this antibody for the forecasting.

As a result of application of the offered multiobjective clonal selection algorithm the Pareto set of the nondominated forecasting models will be received. These models provide the best combinations of values of the used quality indicators for the forecasting models.

IV. Experimental Studies

Both variants of optimization have been applied for the development of the forecasting models intended for forecasting of the names' references' quantity of the E-Commerce systems in the requirements to vacancies posted on the websites of 2 famous recruiting network services – HeadHunter.ru (Russia) and Indeed.com (USA). The obtained forecasting results can be used for the analysis of tendencies of the labour market. Each of the analyzed time series contains information on the number of vacancies which include a specific keyword (Magento, OpenCart, PrestoShop, Hybris, Demandware). This keyword defines the name of E-Commerce system for development of online stores. Herewith 7 TSs have been considered: 4 TSs with information on vacancies in Russia and 3 TSs with information on vacancies in USA marked (in brackets after the name of a keyword) respectively as RF and USA:

Hybris (RF) (monitoring period 03.02.2016 – 06.04.2016, unit of measure is the number of references)=[55; 55; 57; 56; 56; 55; 56; 62; 62; 61; 59; 57; 57; 57; 57; 58; 58; 58; 58; 58; 58; 56; 56; 56; 57; 57; 57; 57; 57; 55; 54; 54; 53; 53; 54; 52; 51; 53; 53; 51; 51; 51; 54; 56; 58; 58; 58; 58; 58; 58; 60; 63; 63; 62; 63; 65; 67; 67; 67; 67; 66; 66; 69; 70];

Magento (RF) (monitoring period 03.02.2016 – 06.04.2016, unit of measure is the number of references)=[116; 128; 130; 132; 131; 126; 125; 120; 128; 130; 133; 130; 128; 124; 121; 120; 123; 123; 127; 129; 123; 127; 129; 131; 133; 133; 123; 123; 123; 123; 123; 120; 120; 118; 116; 117; 118; 129; 124; 120; 120; 127; 130; 131; 129; 129; 124; 124; 129; 133; 134; 132; 133; 133; 132; 135; 136; 136; 136; 136; 136; 142; 140; 147];

OpenCart (RF) (monitoring period 03.02.2016 – 06.04.2016, unit of measure is the number of references)=[84; 84; 82; 81; 80; 81; 89; 88; 89; 85; 85; 85; 88; 91; 93; 90; 88; 85; 90; 90; 91; 85; 83; 83; 77; 77; 77; 79; 79; 76; 72; 75; 73; 74; 76; 80; 75; 80; 80; 79; 86; 92; 93; 89; 90; 90; 90; 92; 94; 96; 98; 99; 99; 101; 115; 116; 116; 119; 120; 120; 112; 114; 114; 118];

PrestoShop (RF) (monitoring period 03.02.2016 – 06.04.2016, unit of measure is the number of references)=[32; 31; 31; 34; 32; 33; 33; 32; 34; 34; 32; 31; 33; 34; 34; 33; 33; 32; 32; 32; 32; 30; 30; 30; 29; 29; 29; 26; 26; 24; 23; 23; 22; 22; 22; 25; 24; 25; 25; 25; 28; 32; 30; 30; 29; 29; 29; 31; 31; 32; 31; 31; 31; 31; 31; 31; 29; 28; 28; 28; 28; 27; 27; 28];

Hybris (USA) (monitoring period 03.02.2016 – 06.04.2016, unit of measure is the number of references)=[674; 688; 677; 672; 664; 680; 690; 693; 710; 706; 697; 692; 688; 689; 676; 668; 663; 648; 643; 637; 635; 642; 641; 641; 629; 629; 629; 631; 631; 666; 671; 663; 667; 672; 711; 703; 662; 715; 715; 709; 690; 665; 668; 657; 662; 659; 659; 660; 657; 658; 656; 653; 653; 646; 643; 630; 631; 649; 647; 647; 652; 654; 662; 650];

Magento (USA) (monitoring period 03.02.2016 – 06.04.2016, unit of measure is the number of references)=[1093; 1102;

1082; 1076; 1076; 1077; 1087; 1095; 1080; 1073; 1072; 1070; 1073; 1086; 1103; 1110; 1118; 1110; 1098; 1107; 1114; 1133; 1126; 1126; 1124; 1124; 1124; 1134; 1134; 1131; 1135; 1125; 1111; 1114; 1138; 1137; 1125; 1155; 1155; 1145; 1103; 1006; 1023; 1013; 1015; 1008; 1008; 1009; 1012; 1021; 1024; 1022; 1022; 1013; 1022; 1023; 1045; 1049; 1040; 1040; 1022; 1018; 1038; 1042];

Demandware (USA) (monitoring period 03.02.2016 – 06.04.2016, unit of measure is the number of references)=[335; 334; 334; 332; 332; 335; 339; 331; 332; 329; 331; 330; 326; 338; 341; 342; 342; 341; 339; 343; 344; 340; 344; 344; 357; 357; 357; 367; 367; 385; 390; 389; 389; 400; 403; 404; 307; 397; 397; 395; 388; 383; 389; 381; 373; 374; 374; 381; 376; 377; 372; 373; 373; 370; 369; 364; 372; 370; 371; 371; 374; 377; 378; 377].

The first 59 values and the last 5 values of elements of each TS were used as the training data sequence and the test data sequence correspondingly. The forecasting models had been developed for each TS with the use of the MCSA on the base of two variants of optimization (Table 1). The forecasting results with use of these models received for the training and test sequences of data are shown in Fig. 3 and 4. The averaged values of the relative forecasting errors at the 5 steps, the averaged values of the affinity indicator and the averaged values of the tendencies discrepancy indicator received by the results of 10 runs of MCSA for each TS are presented in Table 2.

TABLE I. THE FORECASTING MODELS

Time series	The forecasting model on the base of one quality indicator	The forecasting model on the base of two quality indicators
Hybris (RF)	$cos(exp(cos(cos(cos(d(t–4))+cos(d(t–1))) – – sin(d(t–2))))) – exp(ln(d(t–3)))$	$exp(sin(sin(587.749/cos(d(t–2))) – 0.670)+ +sin(d(t–4)+sin(d(t–3))) – sin(0.667 – ln(d(t–1)))) – d(t–1)+cos(d(t–2))$
Magento (RF)	$ln(cos(ln(sin(sin(ln(d(t–1)) – sin(d(t–3)))·d(t–4))+ +cos(d(t–5)) – exp(d(t–2)))/exp(exp(d(t–7))· ·d(t–7)))+exp(d(t–1)+sin(d(t–1))$	$cos(sin(exp(d(t–3)+d(t–2)) – ln(d(t–5)) + +exp(d(t–2)) – sin(d(t–5))+cos(ln(d(t–6))· ·exp(d(t–7))))/ln(sin(d(t–4))+exp(d(t–2))$
OpenCart (RF)	$sin(sin((ln(d(t–2))·cos(d(t–1)))·sin(d(t–5))) – – cos(sin(d(t–6))·d(t–4))·sin(cos(d(t–3))/d(t–3)) – – ln(d(t–3)) – d(t–2))$	$exp(sin(cos(sin(sin(d(t–4))+ln(d(t–3)))+ln(d(t–2)))+ +cos(exp(d(t–6))·ln(d(t–1))) – exp(cos(d(t–3))· ·sin(d(t–2)))) – ln(exp(d(t–1)) – sin(d(t–5))$
PrestoShop (RF)	$ln(sin(exp(sin(ln(d(t–3))·sin(d(t–2)))/ln(d(t–4)))+ +exp(sin(d(t–2))·cos(d(t–5))))·exp(d(t–1)) – – cos(d(t–5))·cos(ln(d(t–1))·sin(d(t–1))$	$sin(ln(sin(sin(cos(d(t–1))·sin(d(t–3))) – sin(d(t–6)))+ +sin(cos(d(t–2))·d(t–4)) – cos(ln(d(t–3)) – – ln(d(t–5)))) – ln(exp(d(t–1))·2.3))$
Demandware (USA)	$ln(exp(sin(cos(d(t–2)) – cos(d(t–3)))·0.002+ +cos(cos(d(t–2)) – d(t–4))·exp(d(t–1) – – sin(d(t–3))))+exp(sin(d(t–2)) – ln(d(t–2))$	$ln(exp(sin(cos(ln(d(t–2))/sin(d(t–4))) – – sin(d(t–3))) – ln(exp(d(t–1)) – ln(d(t–3)))+ +sin(d(t–1)/d(t–3)))+exp(sin(d(t–5)) – ln(d(t–1))$
Magento (USA)	$exp(exp(sin(ln(sin(d(t–6))+sin(d(t–1)))+cos(d(t–5)))· ·cos(sin(d(t–2))·ln(d(t–1))))/exp(exp(d(t–3)) – – cos(d(t–4))))·ln(d(t–1)+cos(d(t–3))$	$exp(sin(sin(sin(sin(ln(d(t–6))·d(t–1))·ln(d(t–3)))+ +cos(sin(d(t–4))/sin(d(t–2))))· ·sin(sin(d(t–5))·0.013))·ln(ln(d(t–4))+d(t))$
Hybris (USA)	$ln(exp(sin(sin(sin(d(t–1))/exp(d(t–3)))·(– 0.499))+ +exp(cos(d(t–4)) – cos(d(t–2)))) – ln(d(t–4)) – – d(t–4))+sin(sin(d(t–1)) – ln(d(t–4))$	$sin(sin(exp((cos(d(t–2)) – 3.530) – cos(d(t–3)))+ +sin(0.968·ln(d(t–4)))·exp(sin(d(t–1))/cos(d(t–5)))) – – d(t–1) – sin(d(t–2)))$

TABLE II. THE AVERAGED VALUES OF THE FORECASTING ERRORS AT THE 5 STEPSAND THE AVERAGED VALUES OF THE QUALITY INDICATORS (AFF (AFER) AND TENDENCY)

№ The name of the TS	*Aff* (*AFER*), %	The value of the forecasting error					*Tendency*		
		1-st step	2-nd step	3-rd step	4-th step	5-th step	Average error of 5 steps, %	for the training sequence	for the test sequence
one quality indicator									
1. Hybris (RF)	2,56	1,12	0,67	0,30	1,55	3,54	1,43	0.45	0.6
2. Magento (RF)	0,44	0,45	0,45	0,45	3,79	2,28	1,48	0.38	0.4
3. OpenCart (RF)	6,86	3,96	3,98	4,27	4,20	4,14	4,11	0.2	0
4. PrestoShop (RF)	1,52	1,32	0,83	0,74	0,55	1,21	0,93	0.09	0.4
5. Demandware (USA)	0,22	0,19	0,20	0,58	0,75	0,44	0,43	0.09	0.2
6. Magento (USA)	2,46	1,30	1,41	0,45	0,20	1,37	0,94	0.43	0.4
7. Hybris (USA)	2,40	3,45	3,45	1,41	1,70	3,07	2,61	0.4	0.4
Average value	2.35	1.68	1.57	1.17	1.82	2.29	1.71	0.27	0.34
two quality indicators									
1. Hybris (RF)	1,06	0,52	0,02	0,02	0,98	0,44	0,39	0.11	0
2. Magento (RF)	0,32	0,27	0,28	0,35	0,48	0,45	0,37	0.06	0
3. OpenCart (RF)	0,32	0,07	0,06	0,12	0,20	0,17	0,12	0	0
4. PrestoShop (RF)	1,04	0,52	0,55	0,44	0,71	0,67	0,58	0.09	0.2
5. Demandware (USA)	0,17	0,15	0,15	0,20	0,23	0,15	0,17	0	0
6. Magento (USA)	0,06	0,05	0,04	0,04	0,09	0,10	0,06	0	0
7. Hybris (USA)	0,08	0,05	0,06	0,05	0,03	0,17	0,07	0	0
Average value	0.43	0.23	0.16	0.17	0.39	0.31	0.25	0.04	0.03

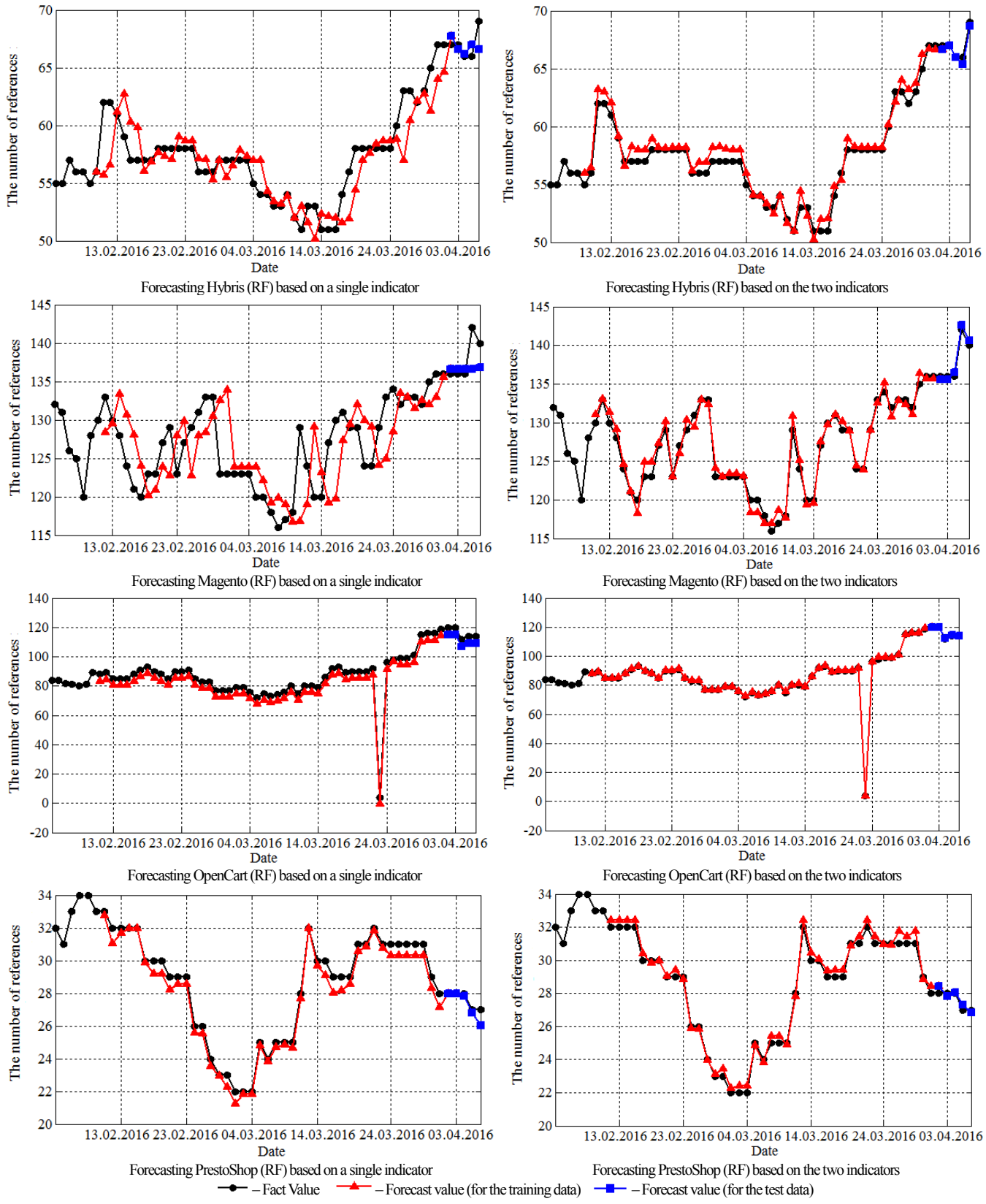


Fig. 3. The Forecasting of TSs, determining the number of references of E-Commerce systems for HeadHunter.ru (Russia)

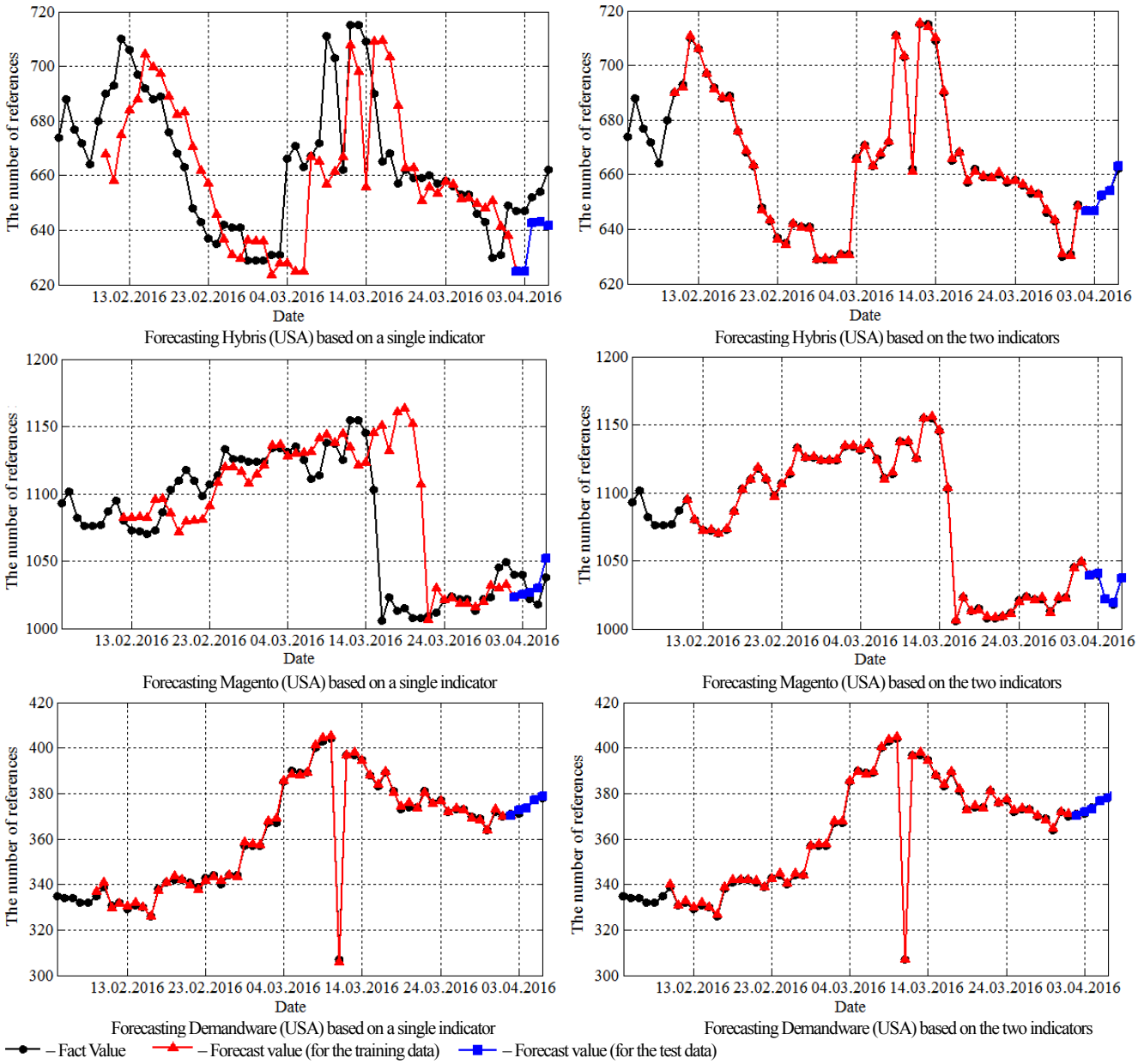
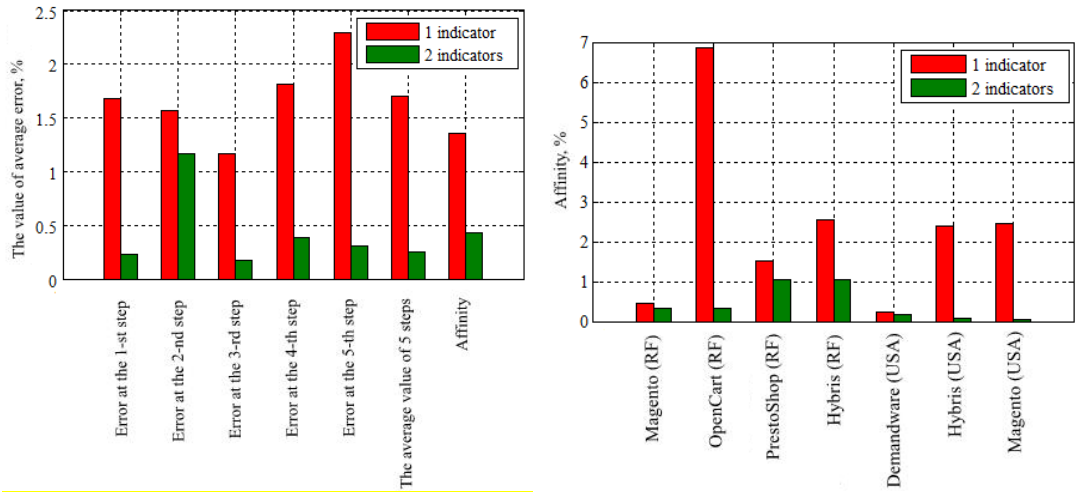


Fig. 4. The Forecasting of TSs, determining the number of references of E-Commerce systems for Indeed.com (USA)

The averaged values of the relative forecasting errors at the 5 steps and the averaged values of the affinity indicator in the context of all TSs are presented graphically in Fig. 5, a. It is clear, that the second optimization variant is more effective as for the solution of problems of short-term forecasting (for 1 – 3 step forward), as for the solution of problems of medium-term forecasting (for 4 and 5 steps forward). Herewith the second optimization variant allows not only receiving the smaller value of the tendencies discrepancy indicator *Tendency* in comparison with the first optimization variant (Table 3 and Fig. 6), but also in many cases reducing the value of the affinity indicator *Aff* (Fig. 5, b) thanks to the corresponding correction of the search direction of the forecasting model.

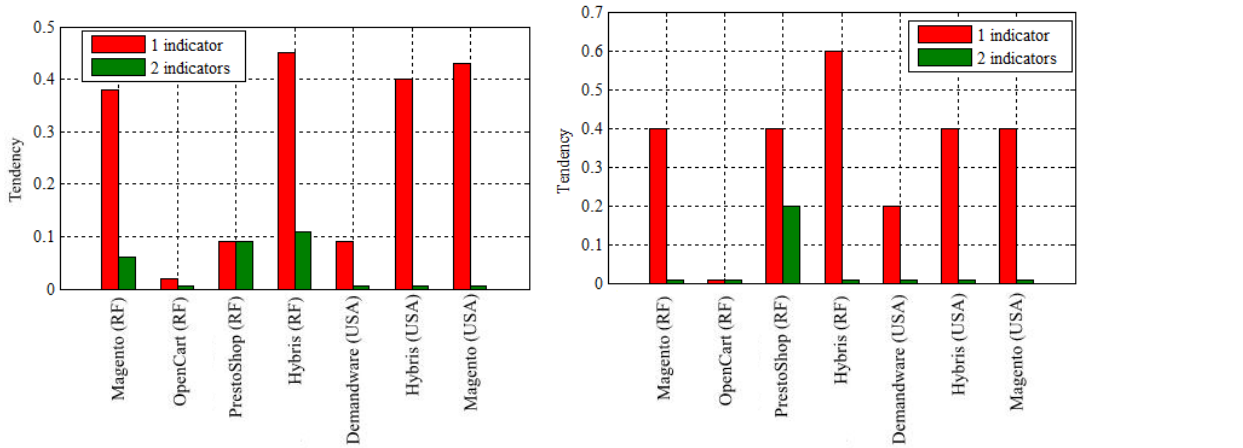
The most essential influence on the development time of the forecasting model on the base of the SBT and MCSA is rendered by such parameters as the number of iterations, size of antibodies' population, coefficient of antibodies' cloning and coefficient of clones' reproduction.



a – The averaged values of the forecasting errors at the 5 steps and the averaged value of the affinity indicator for all TSs

b – The averaged values of the affinity indicator for each TS

Fig. 5. The comparison of the averaged values of the forecast errors at the 5 steps and the averaged values of the affinity indicator for two variants of optimization



a – The averaged values of the tendencies discrepancy indicator for the training data sequence

b – The averaged values of the tendencies discrepancy indicator for the test data sequence

Fig. 6. The comparison of the averaged values of the tendencies discrepancy indicator for two variants of optimization

In the reviewed example 400 iterations of MCSA for population of 20 antibodies were executed. Coefficient of antibodies' cloning was equal to 0.3. Coefficient of clones' reproduction was equal to 0.8. Computer working under the 64-bit Windows 7 version with RAM of 2 Gb and the two-nuclear Pentium 4 processor with a clock frequency of 3.4 GHz was used for experiment. 108.5 seconds on average were spent for creation of one forecasting model on the base of one quality indicator (for the first optimization variant). To build the forecasting model on the base of two quality indicators (for the second optimization variant) it is necessary to spent 120.9 seconds on average that on 12.4 seconds (on 10.3%) more, than for the first optimization variant.

V. Conclusion

The comparative analysis of two optimization variants in the context of the development problem of the forecasting models on the base of the SBT shows the expediency and perspective of use of the second optimization variant which realizes the accounting of two quality indicators of the forecasting model – the affinity indicator and the tendencies discrepancy indicator.

Use of the principles of Pareto-dominance during the MCSA realization at the development of the forecasting models on the base of the SBT allows receiving the effective solution of the accounting problem of two quality indicators of the forecasting model at the acceptable time expenditures. A herewith it is possible to expand the scope of the forecasting models on the base of the SBT and MCSA.

Thus, the expediency of researches on further improvement of the multiobjective optimization algorithms for the purpose of their application to the search problem of the adequate forecasting model of TS is obvious.

REFERENCES

[1] L.A. Demidova, "Time Series Forecasting Models on the Base of Modified Clonal Selection Algorithm", 2014 International Conference on Computer Technologies in Physical and Engineering Applications (ICCTPEA), pp. 33 – 34, 2014.

[2] L.A. Demidova, "Assessment of the quality of the forecasting models based of the strict on binary trees and the modified clonal selection algorithm", Cloud of Science, 1 (2014), pp. 202-222 [in Russian] (http://cloudofscience.ru/sites/default/files/pdf/CoS_2_202.pdf).

[3] N.N. Astakhova, L.A. Demidova, E.V. Nikulchev, "Forecasting Method For Grouped Time Series With The Use Of K-Means Algorithm", Applied Mathematical Sciences, 9, no. 97, pp. 4813–4830, 2015.

[4] N.N. Astakhova, L.A. Demidova, E.V. Nikulchev, "Forecasting Of Time Series' Groups With Application Of Fuzzy C-Mean Algorithm," Contemporary Engineering Sciences, 8, no 35, pp. 1659–1677, 2015.

[5] N. Astakhova, L. Demidova, V. Konev, "The Description Problem Of The Clusters' Centroids", 2015 International Conference "Stability and Control Processes" in Memory of V.I. Zubov (SCP), pp. 448–451, 2015.

[6] N. Astakhova, L. Demidova, "Using of the notion "Pareto set" for development of the forecasting models based on the modified clonal selection algorithm," 6th Seminar on Industrial Control Systems: analysis, modeling and computation, art. 02001, 2016.

[7] C.M. Fonseca, P.J. Fleming, "Multiobjective optimization and multiple constraint handling with evolutionary algorithms – Part I: A unified formulation," Technical report 564, University of Sheffield, Sheffield, UK, pp. 1–16, 1995.

[8] J. Horn, N. Nafpliotis, D.E. Goldberg, "A niched Pareto genetic algorithm for multiobjective optimization," Proceedings of the First IEEE Conference on Evolutionary Computation, 1, Piscataway, pp. 82–87, 1994.

[9] E. Zitzler, L. Thiele, "Multiobjective optimization using evolutionary algorithms – A comparative case study," Parallel Problem Solving From Nature, V, A. E. Eiben, T. Back, M. Schoenauer, and H.-P. Schwefel, Eds. Berlin, Germany: Springer-Verlag, pp. 292–301, 1998.

[10] K. Deb, A. Pratap, S. Agarwal, T. Meyarivan, "A Fast and Elitist Multiobjective Genetic Algorithm: NSGA II, " KanGAL Report No. 200001, Indian Institute of Technology, Kanpur, India, pp. 182–197, 2000.

[11] H. Seada, K. Deb, "U-NSGA-III: A Unified Evolutionary Optimization Procedure for Single, Multiple, and Many Objectives: Proof-of-Principle Results," Evolutionary Multi-Criterion Optimization, 9019, pp. 34–49, 2015.

[12] P. Coello Coello C.A., Cruz Cortés N. "An approach to solve multiobjective optimization problems based on an artificial immune system," Proceedings of the First International Conference on Artificial Immune Systems, University of Kent at Canterbury, UK, pp. 212–221, 2002.

[13] G.-C. Luh, C.-H. Chueh, W.-W. Liu, "MOIA: Multi-Objective Immune Alghorithm," Computers and Structures, 82, pp. 829–844, 2004.

[14] X.L. Wang, M. Mahfouf, "ACSAMO: An Adaptive Multiobjective Optimization Algorithm using the Clonal Selection Principle," 2nd European Symposium on Nature-Inspired Smart Information Systems, pp. 1–12, 2006.

[15] L. Jiao, M. Gong, H. Du, L. Bo, "Multiobjective immune algorithm with nondominated neighbor-based selection," Evolutionary Computation, 16, issue 2, pp. 225–255, 2008.

Big Data Knowledge Mining

Huda Umar Banuqitah , Fathy Eassa,Kamal Jambi, Maysoon Abulkhair
Computer Science
King AbdulAziz University
Jeddah, Saudi Arabia

***Abstract*—Big Data (BD) era has been arrived. The ascent of big data applications where information accumulation has grown beyond the ability of the present programming instrument to catch, manage and process within tolerable short time. The volume is not only the characteristic that defines big data, but also velocity, variety, and value. Many resources contain BD that should be processed. The biomedical research literature is one among many other domains that hides a rich knowledge. MEDLINE is a huge biomedical research database which remain a significantly underutilized source of biological information. Discovering the useful knowledge from such huge corpus leading to many problems related to the type of information such as the related concepts of the domain of texts and the semantic relationship associated with them. In this paper, an agent-based system of two–level for Self-supervised relation extraction from MEDLINE using Unified Medical Language System (UMLS) Knowledgebase, has been proposed . The model uses a Self-supervised Approach for Relation Extraction (RE) by constructing enhanced training examples using information from UMLS with hybrid text features. The model incorporates Apache Spark and HBase BD technologies with multiple data mining and machine learning technique with the Multi Agent System (MAS). The system shows a better result in comparison with the current state of the art and naïve approach in terms of Accuracy, Precision, Recall and F-score.**

Keywords—Knowledge Mining; Relation Extraction; Self-supervised; Big Data; Agent

I. INTRODUCTION AND BACKGROUND

Nowadays large spectrum data is being collected and generated on an unprecedented scale; this paradigm is called "Big Data"(BD)[1].In the last two decades, usage of biomedical computing systems present an explosive growth. The vast amount of Information they store, contains new knowledge that can provide decision support to improve the quality of medical care. MEDLINE is one example of the online bibliographic database on a biomedical domain that contains more than 22 million biomedicine journal articles[2]. As a result, these volumes of data require an efficient prediction and analysis platform to gain fast response and real-time classification for such BD[3]. The ability to discover knowledge from the big data in sufficient time and scalable fashion is a complex task. Data should be processed to extract some helpful knowledge from it. An essential challenge for applications of Big Data is that, the large volumes of data and extracts valuable information or knowledge for future actions[4]. The process of extracting useful knowledge from structured or unstructured data is known as knowledge discovery from Database (KDD) process which refers to a collection of activities designed to obtain new knowledge from complex data dataset[5][6]. KDD from such a biomedical corpus like MEDLINE is a complicated process, and it takes several processes [7]. Information Extraction (IE) techniques are the efficient exploitations of these resources that transform unstructured data into the structured form. An example of these techniques is Relation extraction (RE) which is an automatical mining of relations between the biomedical entities in text. The extraction of the relations between the biomedical entities is the procedure of determining the semantic link between those entities and characterizing the nature of this relationship [2]. Recently RE technigues has found growing interest amongst IE community and many studies concentrate on it because it helps to find new relations and interaction between biomedical entities from raw text and minimize usage of a human resource. RE includes multiple techniques such as Natural Language Processing (NLP), rule–based approach, and Machine Learning (ML) methods[8][9]. There are three types of RE approaches which are: Supervised that uses a corpus of labeled data, Unsupervised method which needs no labeling, and Self-supervised (distant-supervised) that uses a small set of labeled examples. The Unsupervised technique extracts strings of words that exist between the entities in huge amounts of text, and then simplifies and clusters these word strings to produce relation. Unsupervised methods can use massive quantities of data and extract very large numbers of relationships, but the resulting relations may not be simple to map to relations needed for a particular knowledge base.

Supervised relation extraction method, on the other hand, uses ML techniques to solve this problem. This approach requires a sufficiently annotated training data which consists of negative and positive examples. Moreover, the constructing of the annotated data set for training is expensive, time-consuming and requires expert knowledge. Self-supervised approach overcomes this problem by utilizing a knowledge base that includes informations about the exact target relation to automatic annotate the data set. The important assumptions are the sentences contain an entity pairs either represent or not represent a relation will also serve the relationship as well. On the other hand, Self-supervised approaches combine the advantages of supervised approaches, by including the features of noisy pattern in a probabilistic classifier, and Unsupervised methods, by extracting large numbers of relations from big corpora. It is generally believed that in a generic domain, Self- supervision techniques would benefit the relation extraction. However, in biomedical domain, the Self-supervised approach is not perfectly explored yet, becuase of two reasons. The first reason is that in general domain, the Freebase is the basic source of knowledge of Self

supervision technique, which is a lack of biomedical knowledge. The second is, the Self-supervison learning models assume that each entity instance is independent but in biomedical domain, this assumption is violated [10]. Thus a system model for Self-supervised Relation Extraction from Biomedical domain was proposed. As mentioned previously, KDD is iterative and interactive multiphase processes that include different steps like the selection of data , preparation and preprocessing, transformation of data , Data Mining (DM) and evaluation process. DM is the core process of KDD, and many researchers interested to integrate between DM and agents. DM can take benefit from agent through involving the intelligence to data mining system while the agents can take benefit from data mining through extending knowledge discovery capability of agents. There is some application that designs the process of KDD assimilates some modules to an agent, they proposed a strategy for integrating different techniques for mining database from agent perspectives [7]. For that, every module of the system was assigned to an agent to get the benefit of the agent technology in data mining process and improve overall system performance. diverse techniques for data mining have been integrated including Self-Supervision, natural language processing, machine learning and Multi-Agent System (MAS) to build a generalized Relation Extraction system from MEDLINE that requires minimal supervision using Unified Medical Language system (UMLS[1]).

The aim of this paper is to develop an agent based knowledge discovery system model for Self-supervised relation extraction in MEDLINE biomedical domain using UMLS knowledge base. Additionaly, different text features were implented with a paragraph to vector model and evaluate by using different classification algorithm to demonstrate the best algorithm with best features that can enhance the model performance for relation extraction. In addtion, Spark[2] and HBase[3] BD technology are integrated to speeding up the processing and accesing of such BD. The model has distinct two characteristics that distinguish the work from the existing ones which are: first, the construction of training example by using the semantic type of the concepts pair in MRREL section of UMLS is a new method of the exciting works in supervised relation extraction in the biomedical domain, and the second is using paragraph to vector model that transfer the sentence to vectors and using the resulted vectors as additional features with other features to improve the classifier performance; these characteristics improves the result comparing with others in terms of Accuracy, Precision, Recall and F-Score performance.

The rest of the paper organized as the follow. Section II presents the related work of the study, while section III describes the details of system architecture. Section IV introduces the methods used in the two levels of the proposed system and the experimental setup with the used dataset. Section V shows the results and the discussion while the final section is the conclusion.

II. Related Work

This section presents the different efforts that have been achieved in relation extraction from a biomedical domain which using distance supervised approach.

The author in [11] represents The general distant supervision approach for relationship extraction as following.

1) Identify a knowledge base which includes pairs of entities about the relationship-type in question (e.g., PPI-database).

2) Compile a large text (not annotated) resource relevant for the target domain (e.g., MEDLINE abstracts).

3) Recognize and normalize relevant named entities (e.g., protein names).

4) Associate entity-pairs from the knowledge base with previously identified instances in the text corpus.

5) Entity pairs contained in the knowledge base are labeled as positive instances. Negative instances are labeled by following the closed world assumption. The closed world assumption states that entity pairs lacking in the knowledge base do not feature the relationship type in question.

There are limited works which used Self-Supervised approaches in the biomedical domain. Most of these papers have used only the abstract of each paper, by utilizing the coordination structure of an entity in the sentences, [10] built up a Self-Supervised model which consolidates the result from open data extraction methodologies, to implement a task of relation extraction from biomedical resarch paper. They consider the structure coordination among entities that co-occurred in one sentence, is done by incorporate a grouping strategy to their model. They apply the Self-supervision technique to extract relationship of gene expression between genes and brain regions from literature. The Results showed that the model accomplish a better performance using Support Vector Machine (SVM) and with non-grouping strategy.

In [12] the authors trained the classifier using Self-supervision technique for Protein-Protein Interactions (PPI). They use a SVM classification algorithm as a classifier. IntAct database is the source of knowledge about interacting proteins.

Using UMLS as Knowledgebase, the authors in [13] proposed a Self-supervised approach for relation extraction from biomedical domain in MEDLINE abstracts using UMLS to annotate automatically the training data which is then used to train the classifier. To generate the training examples with positive and negative examples, all Concept Unique Identifier (CUI) pairs for the target relation are taken from MRREL and consider as a set of positive pairs. Hence, the presence of positive pair entities in a sentence will represent the target relationship. Any sets which additionally happen in another MRREL relations are expelled from the list of positive examples set. Conversely, negative instance will be detected depending on the positive pairs; new CUI pair combination will be created by joining all CUIs from the first position with all CUIs from the second position. These new combination will considered as a negative instance pair, only if a newly

[1] Unified Medical Language System (UMLS)
[2] http://spark.apache.org/
[3] http://hbase.apache.org/

produced CUI pair is not in the positive list and not appear in another MRREL relation. The model evaluated using two techniques Held-out and manual evaluation. On manual evaluation, the classifier was trained using the relation (may_treat), that created using Self-supervised and evaluated by using manually annotated corpus using test data set, and the result outperforms naïve approach with an F-Score of 0.571, 0.600 Precision and 0.545 Recall. The result indicated that UMLS is a useful resource for Self-supervised relation extraction. Additionally by utilizing UMLS to training a Self-supervised relation classifier,[14] exhibited the primary results utilizing UMLS knowledge base and the model assessed by utilizing the existing data set, since there were no directly annotated resources with UMLS relations is available. The presented model in [14] determined that utilizing a Self-supervised classifier which trained on MRREL relations like those found in the evaluation data set, will give propitious results.

The authors in [15] demonstrated the potential of Self-supervised learning in constructing a fully automated relation extraction process. They produced two distantly labeled corpora for drug to drug and protein–protein interaction extraction, with knowledge found in IntAct database for genes and Drug Bank database for drugs. They labeled approximately 50,000 MEDLINE abstracts using the shallow linguistic classifier trained on a distantly labeled corpus. In other words, the classifier trained on five manually annotated corpora and the same classifier trained on a distantly labeled corpus agree on 86.4 % of all 50,000 predictions.

There are some works done in Sel-Supervised approach outside the biomedical domain. Mintz and others in [16] provide relation extraction using Freebase for Self supervision. They utilized the same heuristic by matching tuples of Freebase with unstructured sentences from the Wikipedia articles in their experiments to produce features for learning relation extractors. instead of matching Wikipedia infobox with corresponding Wikipedia articles, matching Freebase with arbitrary sentences will potentially increase the size of matched sentences at the cost of accuracy. They conclude that their results suggest that syntactic features are quite useful in Self-supervised relation extraction. Also, the authors in [17] used Freebase knowledge base to annotate the corpus of New York Times with pairs of entity. They concentrated on the three basic relations which are birth place, nationality and contains. To prepare the classifier for training, they presented the utilization of a multi-instance learning technique for this context. In contrast, the authors in [18] annotated the information in the articles of Wikipedia using the infoboxes of Wikipedia as a knowledge source.

III. System Architecture

The system model consists of two main levels each with its own agents as shown in Fig1. The next subsections describe in details the components, functionality and the implementation of each level.

A. Level 1

The First level deals with data preparation and extraction, relation labeling with the usage of UMLS knowledge base, features extraction and training classifier on resulting train set.

1) Data preparation and extraction

MEDLINE corpus[4] is used as initial data. Medline is a large corpus of biomedical abstracts and articles. The sentences of MEDLINE contain the information of interest such as the biomedical entities. To use MEDLINE for the proposed Self-supervised system model, it should be annotated with these entities. And since UMLS KB was used to construct the training example in Self-supervised approach. So a mapping of UMLS concepts to the MEDLINE sentences is needed. For that, we used a MetaMapped MEDLINE, which is annotated by MetaMap tool[5]. Each sentence in MEDLINE annotated with UMLS concepts, and the annotations are represented in MetaMap machine output format[6].

UMLS is a collection of software and files that incorporate diverse biomedical knowledge base and vocabularies. Metathesasurus is a database in UMLS and contains a huge number of health and biomedical-related concepts and names and the relationship between them. all concepts arranged by their semantic type and all concept names are unified by Concept Unique Identifier (CUI). MRREL[7] form a small part of the Metathesasurus and includes diverse relations between various biomedical concepts which characterized by a couple of CUIs.

By following [13], tables from Metathesaurus have been used, which contains a mapping from Concept Unique Identifier (CUI) to Type Unique Identifier (TUI). MRREL defines binary relations between concepts, for that, a specific relations were used, these relations identified with "RO" keyword such as "may treat," "may prevent" and "gene product malfunction associated with disease". Those relations are most common for relation extraction task. Also, two semantic types have been used which are "**bacs**" and "**dsyn**" pairs, where "bacs" is Biologically Active Substance and "dsyn" is refer to Disease or Syndrome.

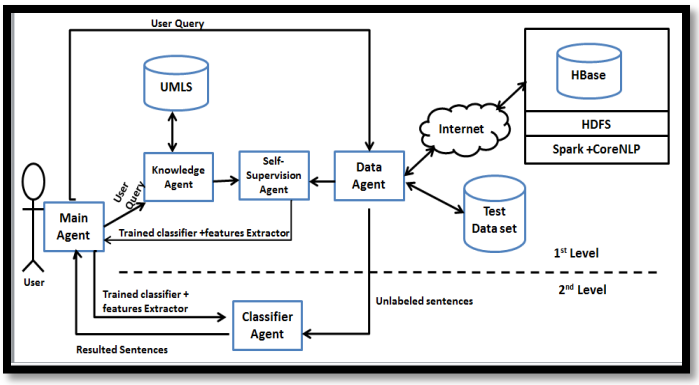


Fig. 1. System Architecture

[4] http://ii.nlm.nih.gov/MMBaseline/
[5] https://metamap.nlm.nih.gov
[6] https://metamap.nlm.nih.gov/Docs/2012_MMO.pdf
[7] MRREL table description

In the system framework, the main query to the knowledge base is taking all the relations between given semantic types. To make execution of the query fast, different tables from Metathesaurus have been joined to have all the needed information, so the resulting table contains pair (CUI1), (CUI2), relation and (TUI) of concepts.

The most time-consuming part in the system is getting sentences from Medline corpus that match the user query. Since the data size is too big to be handled by single commodity machine, advice from [6] was followed and stored these BD inside Hadoop distributed file system.

The main type of query in the system is getting all the sentences from Medline that has entities of a user-specified pair of UMLS semantic types. HBase is the only database that meets all the requirements. HBase is used to handle a large amount of data. It is designed to perform a fast linear scan on large collections, which can be used to perform fast queries.

To get the corpus in HBase, the files were located in HDFS file system. Then Spark workers have been run. Each worker takes a separate file and performs parsing, CoreNLP processing, and conversion to JSON format of each sentence. Hadoop and Spark help to do this job highly parallelizable – each small file can be processed independently.

2) Features Extraction

In the system model, the text are represented in multiple features to train the classifier by using two models, Bag of word model and paragraph to vector model. The description of each model with features will be in the following paragraphs:

a) Bag of world model features

Bag of world model is simple representation used in NLP. In this model the text such as sentences or document is represented as the bag of its words, disregarding of word semantic meaning or ordering in the text. The same text features were adopted, that depend on this model and implemented by [14] [16] [19] because they clearly represent the relation between the entities in the sentence also they help in determining the accurate class of the relation between disease and treatment. The adopted features are the sequence of words between entities, Post Of Speech tag (POS) of words between entities, Words on the semantic path between entities. For constructing these lexical and syntactic features,each sentence was annotated in the training set with part of speech tags and dependency tree using Stanford CoreNLP library[20], which has a large variety of instruments including parser, lemmatizer, tokenizer, part of speech tagger and it is written in Java programming language, which makes it easy to use inside Hadoop ecosystem.

a) Paragraph to vector model features

A paragraph to vector, or called (Doc2vec), is a method for constructing distributed vector representation for sentences and text documents [21]. In fact, the model "Doc2vec" has the potential to overcome many weaknesses of "bag-of-words" model. First, they inherit the most importan property of the word vectors: the semantics meaning of the words. The second advantage is that they take into consideration the word order considering word order and mapping semantically close words to close vectors. It was experimentally shown, that paragraph vectors can be better than other features for document classification task, this because the important characteristic of paragraph vectors is that they are learned from unlabeled data and thus can work well for the tasks that do not have enough labeled data. The paragraph or sentences in the model are mapped to a unique vector that can be used as features for the sentences; then these features can feed directly to conventional machine learning techniques such as logistic regression, support vector machines or others[21][22]. Therefore, from the previous characteristic of Doc2vec model, the model in [23] have been used to add the resulted vectors of sentences as addition features to represent the sentences and how the entities in the sentence related to each other, this to improve the relation extraction model.

3) Level 1 Agent Functionality

a) The main agent

interacts with system user and coordinates other agents

b) Knowledge agent

retrieves relations that correspond to a user query. Relations are represented as a triplet (CUI1, relation name, CUI2). Current knowledge agent implementation uses MRREL as a source of relations.

c) Data agent

finds sentence objects that correspond to a user query. Each sentence object contains the following information:

1) Id – unique identifier of a sentence, for Medline sentences it contains paper id.

2) Text – text of a sentence

3) Mappings – mappings from word to medical entity provided by Metamap tool. Mappings contain information about the semantic type, CUI, name and position in the text of matched entity.

4) Tokens – representation of CoreNLP parse results. Each token contains information about its POS, head, dependency relation, lemma and position in the text.

Data agent operation is chunk-based. When data agent receives a query, it gets result in a small chunk and transmits them to another agent one by one which helps to reduce total query time because other agents can start their work early.

a) Self-supervision agent

This agent constructs relation classifier with Self-supervision method. Self- supervision agent trains classifier. It constructs labeled training set without human intervention.

Self-supervision agent depends on knowledge and data agents. Knowledge agent provides relation examples, and data agent provides sentences and their parse results.

Self- supervision agent operates in several steps:

1) Constructs train set by matching sentence objects from data agent and relations from knowledge agent. This job is done chunk-wise. The result of this work is automatically labeled train set.

2) Fits feature extractors on train set and extracts features.

3) Trains classifier on extracted features.

4) Returns trained feature extractors and classifier to main agent.

Note that all relations are divided into two groups – 'general' and 'specific'. General relations most commonly are synonymous or 'is-a' relation. Specific relations represent more complex interactions between entities, for example, 'may treat' or 'may prevent'. Specific relations has label 'RO' in UMLS.

The following steps are the labeling train set algorithm:

1. For each sentence:
   a. Get all relations that have CUI1 and CUI2 same as CUI's of sentence entities
   b. If all matched relations are general – label sentence as negative example ("other")
   c. Else if sentence matched several specific relations or matched no relations – filter it out
   d. If sentence matched single specific relation and none of general relations – label it with specific relation
2. If some relations represent less than 5 % of train set – filter them out.

After labeling, Self- supervision agent performs feature extraction and classifier training. Then these extractors and classifier are sent to main agent.

B. Level 2

The second Level applies trained classifier to new data to label the unlabeled sentences.

1) Level 2 Agent Functionality

a) Classifier agent

This agent receives trained feature extractors and classifier from main agent. Then it gets an unlabeled sentence in chunks from data agent. For each chunk, it extracts features and performs label prediction with the classifier. Labels with sentence text and id are returned to the main agent. In the experiment different classification algorithms were used, including k-Nearest Neighbors, Linear SVC, and logistic regression.

IV. Experiments

To evaluate the proposed system, the system model was compared with the proposed system in[13]. This done by constructing training data set and two tests set.

A. Tools

Experiments were conducted in InteliJ IDEA which is integrated development environment (IDE) for Java because its maximize developer productivity. For agent system, JADE framework was used as a most contemporary and well-documented agent-based framework.

As mentioned before, Hadoop ecosystem and HBase have been used for BD storage. Stanford CoreNLP library was used for data preprocessing using Apache Spark. LIBLINEAR library is used for classification and evaluation metrics. For doc2vec model, GENSIM library was used.

B. Agent system implementation

Agent abstraction was incorporated in the framework of the system because it makes easier to build extensible distributed systems with a lot of communicating entities. JADE framework [24][25] was used as a most contemporary and well-documented agent-based framework. In addition to agent abstraction, it provides built-in task composition model, peer-to-peer communication, and agent subscription service.

As shown in Fig1, the system consists of following entities: agents, data models, feature extractors, and classifiers. Agents are the ancestors of JADE's agent class and represent independent steps of the knowledge discovery pipeline.

In addition, a Main Agent coordinates pipeline execution and manages the User Interface (UI). Despite of this, other agents can operate independently. For example, one can query knowledge agent for available relations.

Feature extractors and classifiers represent different features and classification algorithms used for relation extraction. Agents communicate via JADE messaging system and JADE yellow pages. There are two types of communication messages, coordination messages, and payload messages. Coordination messages serve to orchestrate user query execution among agents. Payload agents carry data such as sentences or classifiers. Jade yellow pages were used by the Main agent to check and get the list of agents who exist on the system

C. Features extraction

For a bag of word model features, feature-specific information has been extracted from the train sentences in the form of a token set. Then Term Frequency-Invers Document Frequency (TF-IDF) algorithm was applied. If several features were used for classification, resulting feature matrix is obtained by concatenation of feature vectors for both features.

For doc2vec, the recommendations of [23] article was followed. both distributed bag of words and distributed memory variations of the algorithm have been used.

For all the classification algorithms, the default parameters and settings were used.

D. Training set construction

The training set was constructed from sentences that matched MRREL relations with our own method as mentioned in section 3 and inspired by[13]. However, the model differs in two aspects: a semantic type of the entities is used to get all the relations between the biomedical entities in UMLS KB, and we used general relation examples that appear between the given semantic types to construct the negative examples. In contrast authors of [13], used only pairs that participate in "may_treat" relation, regardless of their semantic type.

To enhance the training set quality, we applied filtering by part of speech tag. MetaMap tool has a most common error that is annotating verbs or adjectives as if they were nouns as observed by manual check. Using CoreNLP library as in [20] we annotated each sentence in training set with part of speech

tags and threw away those sentences which concept was not marked as nouns.

For the training set labeling, all relations were divided into two groups: specific relations that labeled with "RO" in MRREL, where RO relation described as has a relationship other than synonymous, narrower, or broader, and other than RO relation groups that represent more general relations. General relations were considered as negative examples for classification and labeled as "other." Sentences with multiple "RO" relations were not included in a training set because they could represent any of those relations, but classifier needs the exact match with label and ground truth. We also discard non-frequent relations.

Another observation was that "RO may_treat" relation almost include"RO=may_prevent" relation and all most of the sentences labeled with "may_prevent" were also labeled with "may_treat". Manual analysis showed that ground truth for such sentence could be either of both relations as shown in example 1 that the treatment "desferrioxamine" treats the" iron overload", and they are indistinguishable by MRREL. We decided to unite such relations into one more general.

Example 1: [Intensified **desferrioxamine** (TREATMENT) treatment (by either subcutaneous or intravenous route) or use of other oral iron chelators, or both, remains the established treatment to reverse cardiac dysfunction due to **iron overload** (DISEASE)]

Since our target examples of relation is "may_treat" we observed that "null" and "related_to" relations will not serve this relation between treatment and disease entities, if we consider example 2, we can observe that "METABOLIC SYNDROME" does not treat or prevent the disease "CHOLESTEROL", but they are related to each other in another way. For that, we exclude "null" and "related_to" examples from the training data set examples.

Example 2: [BACKGROUND: To establish the rate of agreement in predicting **METABOLIC SYNDROME** (TREATMENT) (ms) in different pediatric classifications using percentiles or fixed cut-offs, as well as exploring the influence of **CHOLESTEROL** (DISEASE)]

E. Test set construction

Two data sets have been used to evaluate the performance of the classifier model. The first test set constructed by combining different relation mining data sets so that it could be similar to a training set. The second test set we used the same test set presented in [13] after their permission.

In the first test set, we employed three most specific and frequent relations: "may_treat", "gene_product_malfunction_associated_with_disease " and "other" to serve our training set that contains these relations. Further, we identify this data set as "Triple relation" test set (for simplicity). For this test set, 70 examples of "other" relation were labeled manually. 500 "may_treat" examples and 60 "other" examples were obtained from disease-treatment relations test set in [19]. 500 examples of "gene_product_malfunction_associated_with_disease" were randomly chosen among positive examples of gene-disease relation test set in [26].

The second test set from [13] contains 227 examples of "other" relations and 173 examples of "may_treat" relations. This set is called "may_treat." test set. Since it is important to keep in training set only those relations that presented in the test set, we exclude the relation "gene_malfunction_is_associated_with_disease" from the training set examples to evaluate using the test set "may_treat" from [13].

V. Result and Discussion

Because preprocessing works independently for each sentence, this job is highly parallelizable. We used Spark framework to do the parallelization. Since data was represented as a collection of compressed files, parallel processing was done file-wise. Performance results are summarized in Table. 1 when using a Spark in preprocessing step with CoreNLP for 4000 sentences, which indicate that using spark with a different number of worker reduce the time which means it speed up the processing step.

Different measurements are used to measure the performance of the system. The main purpose of measuring the performance is to compare the system with other systems to determine the success of the proposed design. In the literature, the most widely used evaluation metrics are Accuracy, Precision, Recall, and F-Score. Thus we used these measurements that most common metrics used in classifier evaluation which defined in equations (1), (2), (3) and (4) respectively:

$$Accuracy = \frac{tp+tn}{tp+tn+fn+fp} \quad (1)$$

$$Precision = \frac{tp}{tp+fp} \quad (2)$$

$$Recall = \frac{tp}{tp+fn} \quad (3)$$

$$F - Score = 2 * \frac{precision*recall}{precision+recall} \quad (4)$$

Where (tp) is the true positive results of classification and (fp) is the false positive results of classification and (fn) is the false negative.

On "Triple test set", the values of Precision, Recall and F-Sore has calculated for each class, and then a weighted average is calculated.

TABLE I. PERFORMANCE OF MEDLINE DATA PREPROCESSING USING SPARK FRAMEWORK

Experiment number	1 worker	2 workers	4 workers	No spark
1	239.9	161.1	146.8	434
2	245.9	163.2	144.4	420
3	245.2	162.8	143.8	411
4	241.4	165.4	144.2	420
5	244.9	170.1	143.6	434
Average	243.46	164.52	144.56	423.8

Based on Fig 2, the best result of self-supervised approach in [13], achieved when the baseline data set restricted to 10,000 training instances.

On the system, a different combination of features that discussed in section 3 and different classification algorithm have been applied to evaluate the model on both test sets. Based on "Triple test set", the model shows a better result in terms of Accuracy and Precision when using Linear SVM as the algorithm of classification and Words between entities as basic feature to represent the sentences as shown in Fig. 3 and in term of Recall, and F-Score by using KNN with Euclidean cosine metric with words between entities and words on semantic features. On the same test set and by applying paragraph to vector as an additional feature with the other features as shown in Fig. 4, the best result achieved in Precision, Recall, and F-Score, when using Linear SVM with a paragraph to vector concatenated with words on the semantic path and words between entities features, and in term of Accuracy by using Logistic regression with paragraph to vector concatenated with words on the semantic path and words between entities features.

Furthermore and based on "may_treat" and in comparison with paper[13], the better result as shown in Fig. 5, achieved in the term of Recall and F-Score when using Words between entities features with words on semantic path features using Linear SVM algorithm and in terms of Accuracy and Precision when using Logistic regression with words between entities feature. Fig. 6 shows that the best result after adding a paragraph to vector as an additional feature with the other features is achieved in term of Recall and F-Score by concatenating paragraph vectors with words on the semantic path and words between entities using Linear SVM algorithm. In term of Precision the best result achieved by using KNN with cosine distance metric and paragraph vectors with words between entities and words on semantic path features. The best Accuracy result achieved by applying paragraph to vectors with words on the semantic path using Logistic regression.

The above discussion showed that the system results outperform results from [13]. The reason is that the authors in [13] took sentences that contain random disease-treatment entity pairs which not presented in knowledge base in "may_treat" relation, but due to incompleteness of actual UMLS MRREL knowledge base, those pairs are still very likely to have the target relation "may_treat", so they will have some portion of positive sentence labeled as negative that confuse classifier and harm its performance. On the other hand, in proposed method, only the pairs, which participate in relations other that the target relation, was used to label the negative examples. Those pairs are much less likely to be positive examples, so the train set has higher labeling quality which increases classifier performance. Also, this hypothesis was also confirmed by visual analysis of obtained train sets.

Moreover, by using paragraph vector features, the system results increased as shown in Fig. 4 and Fig. 6, the reason, as justified in [21] and[22], is that in contrast to other features of a bag of the world model, doc2vec captures word semantics that gives additional information to a classifier which enhances the classifier performance.

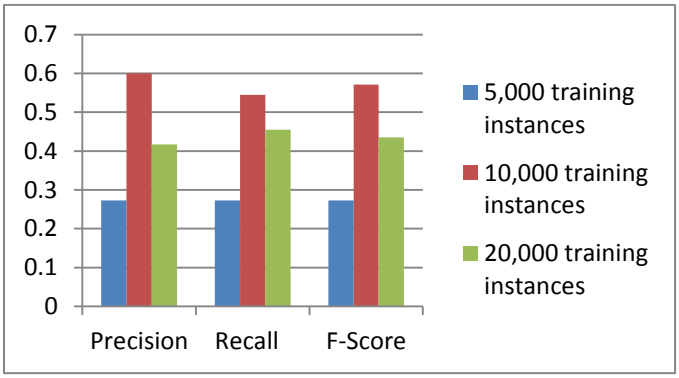


Fig. 2. The result of [14] based on "may_treat" test set

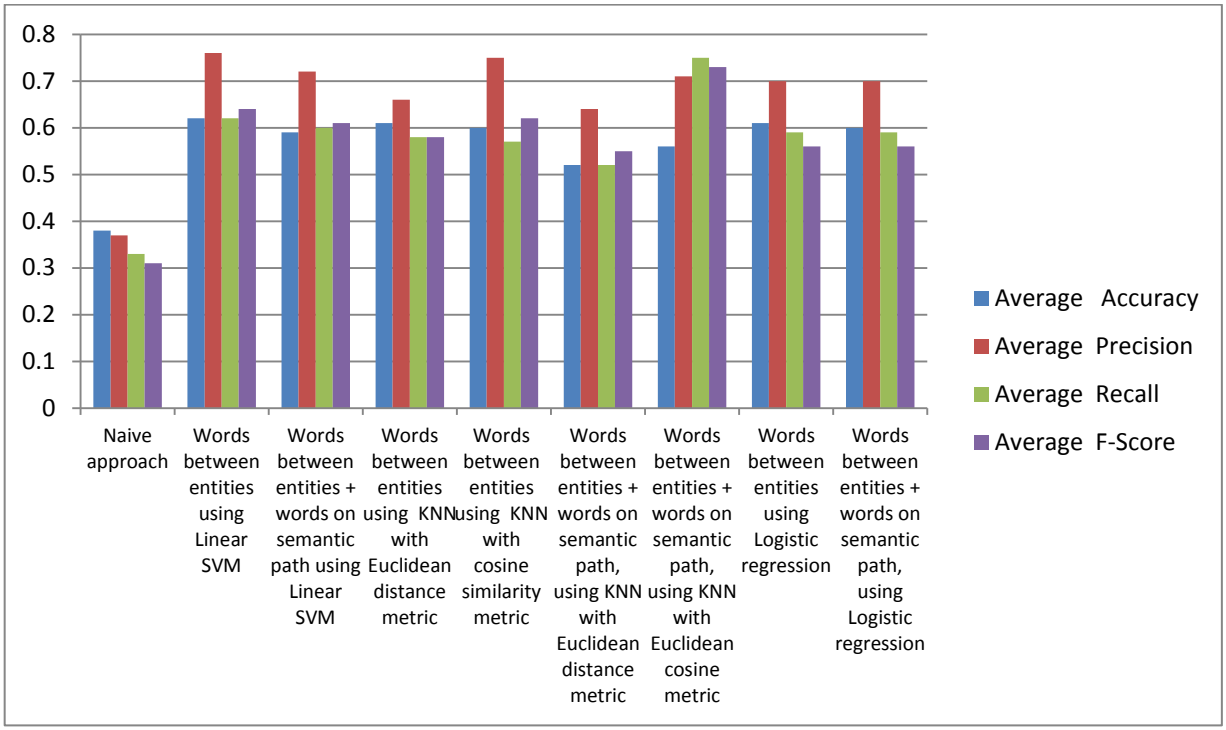


Fig. 3. The result of "Triple test set"

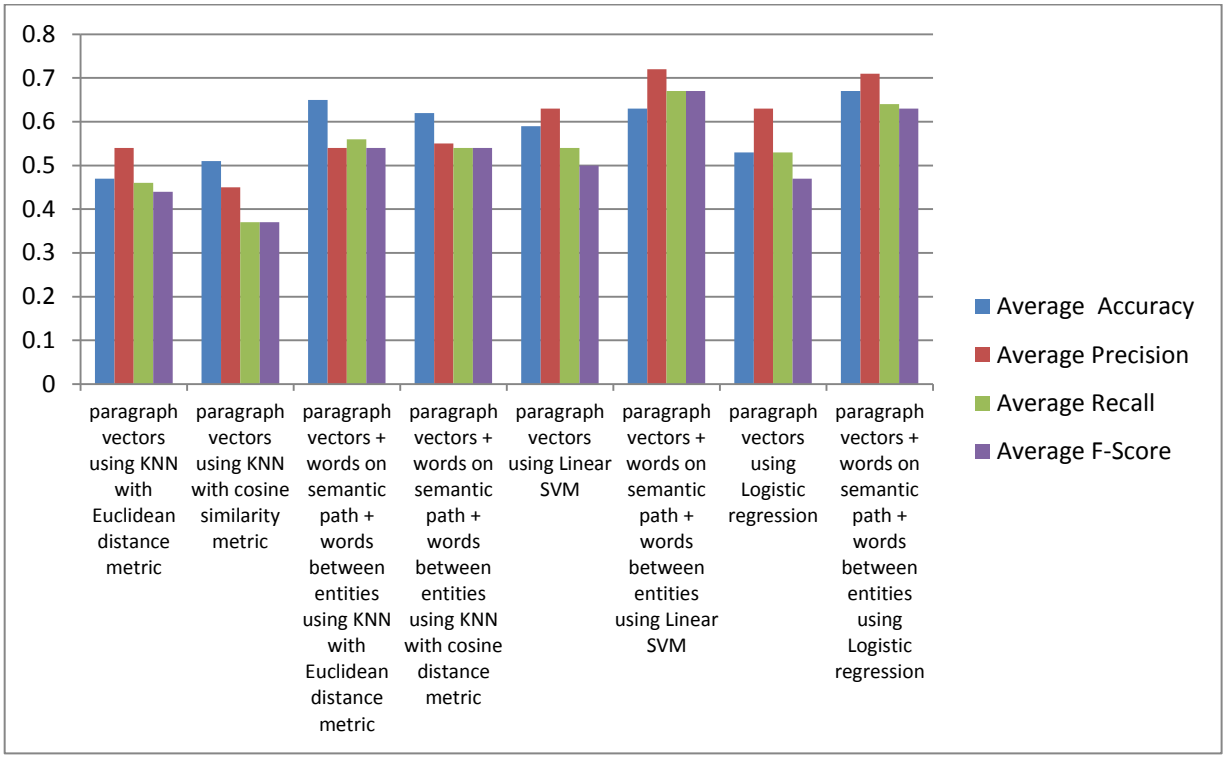


Fig. 4. The result of "Triple test set" with paragrph to vector

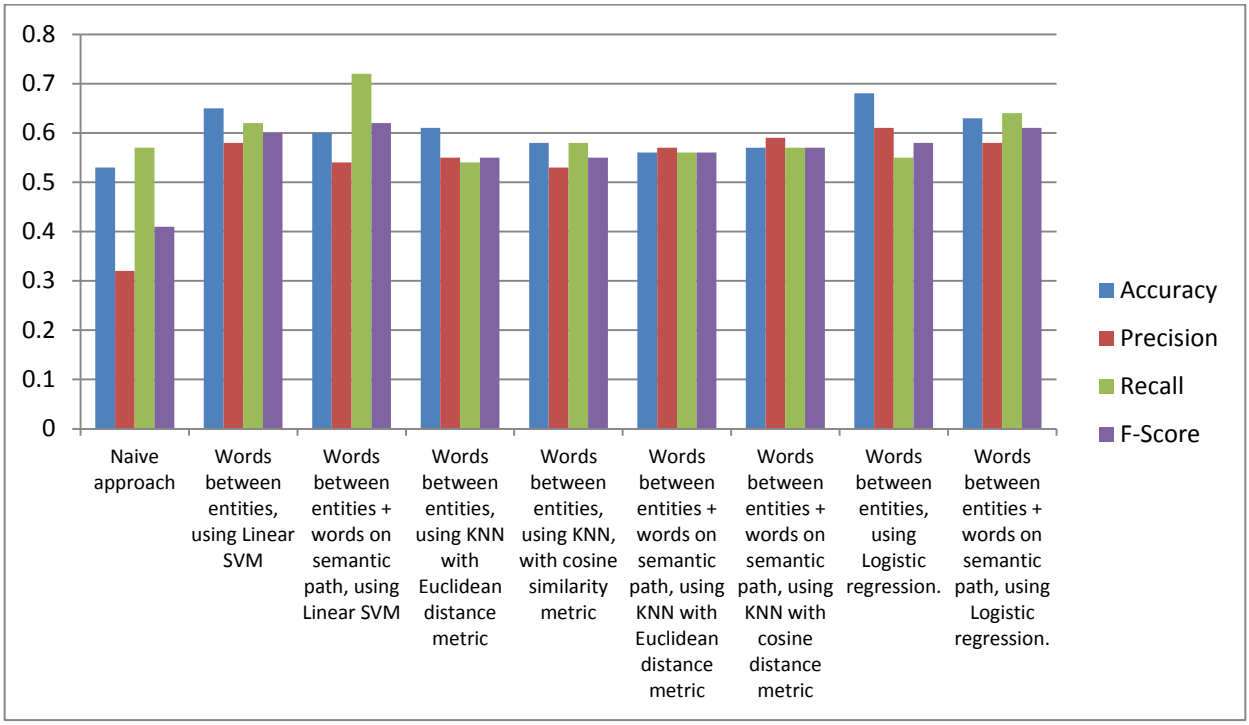


Fig. 5. The result of "may_treat" test set

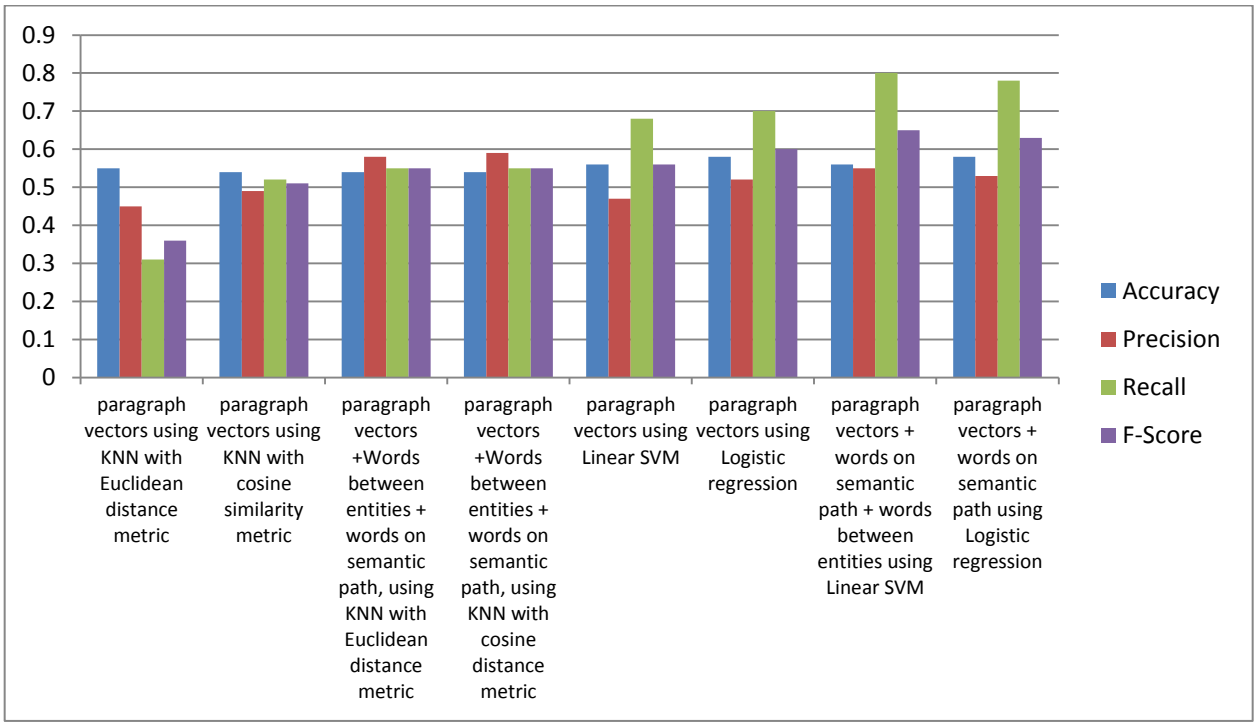


Fig. 6. The result of "may_treat" test set with pargraph to vector

VI. Conclusion

This paper presented a model of big data knowledge mining for Self-supervised relation extraction between biomedical entities from MEDLINE biomedical texts using UMLS knowledge base. The system model fundamentally focused on the extraction of semantic relations between treatments and diseases. The model use hybrid features sets which are: The document to vector, sequence of words between entities, words on the semantic path between entities to enhance the classification performance. The system use a Self-supervised approach for relation extraction by incorporating DM and ML with MAS techniques and demonstrate model performance on MEDLINE data and

UMLS knowledge base to constructing training examples. Moreover, the model used Spark technology with HBase to speeding up the processing of such BD corpus which indicates that using spark with more than two workers will speeding up the preprocessing step. The results also showed that the presented model achieved better results by adopting different features representation and running different classifier algorithms comparing with outperform naïve approach and other paper approach in terms of Accuracy, Precisions, Recall and F-Score. The model also demonstrates an approach to minimize the cost of relation extraction by using a weekly labeled training example using UMLS.

VII. Scope of Future Work

The future work can be classified into two categories, first: improving the performance of relation extraction quality by using Bootstrapping relabeling technique in [27], which can enhance labeling quality. Second improvement is extending train set size with usage of several knowledge bases such as UMLS and IntAct. Using both contain examples of protein-protein interactions. Using both of them can gather more examples and label more data and building manual mapping to unify all the relation representation of each knowledge base or develop sophisticated algorithm that can discover such mapping automatically.

References

[1] M. R. Wigan and R. Clarke, "Big Data's Big Unintended Consequences," *Computer*, vol. 46, pp. 46-53, 2013.

[2] A. Bchir and W. Ben Abdessalem Karaa, "Extraction of drug-disease relations from MEDLINE abstracts," in *Computer and Information Technology (WCCIT), 2013 World Congress on*, 2013, pp.3-1 .

[3] W. Xindong, Z. Xingquan, W. Gong-Qing, and D. Wei, "Data mining with big data," *Knowledge and Data Engineering, IEEE Transactions on*, vol. 26 ,pp. 97-107, 2014.

[4] L. R. Sebastian, S. Babu, and J. J. Kizhakkethottam, "Challenges with big data mining: A review," in *Soft-Computing and Networks Security (ICSNS), 2015 International Conference on*, 2015, pp. 1-4.

[5] O. Rusu, I. Halcu, O. Grigoriu, G. Neculoiu, V .Sandulescu, M. Marinescu, *et al.*, "Converting unstructured and semi-structured data into knowledge," in *Roedunet International Conference (RoEduNet), 2013 11th*, 2013, pp. 1-4.

[6] E. Begoli and J. Horey, "Design Principles for Effective Knowledge Discovery from Big Data," in Software Architecture (WICSA) and European Conference on Software Architecture (ECSA), 2012 Joint Working IEEE/IFIP Conference on, 2012, pp. 215-218.

[7] S. Benomrane, M. Ben Ayed, and A. M. Alimi, "An agent-based Knowledge Discovery from Databases applied in healthcare domain," in *Advanced Logistics and Transport (ICALT), 2013 International Conference on*, 2013, pp. 176-180.

[8] V. N. Romero, S. Kudama, and R. Berlanga Llavori, "Towards the Discovery of Semantic Relations in Large Biomedical Annotated Corpora," in *Database and Expert Systems Applications (DEXA), 2011 22nd International Workshop on*, 2011, pp. 465-469.

[9] Y. Lin, S. Cheng-Jie, W. Xiao-Long, and W. Xuan, "Relationship extraction from biomedical literature using Maximum Entropy based on rich features," in *Machine Learning and Cybernetics (ICMLC), 2010 International Conference on*, 2010, pp. 3358-3361.

[10] L. Mengwen, L. Yuan, A. Yuan, H. Xiaohua, A. Yagoda, and R. Misra, "Relation extraction from biomedical literature with minimal supervision and grouping strategy," in *Bioinformatics and Biomedicine (BIBM), 2014 IEEE International Conference on*, 2014, pp. 444-449.

[11] P. Thomas, "Robust relationship extraction in the biomedical domain," Mathematisch-Naturwissenschaftliche Fakultät, 2015.

[12] P. Thomas, I. Solt, R. Klinger, and U. Leser, "Learning protein protein interaction extraction using distant supervision," *Robust Unsupervised and Semi-Supervised Methods in Natural Language Processing*, pp. 34-41, 2011.

[13] R. Roller and M. Stevenson, "Self-supervised Relation Extraction Using UMLS," in *Information Access Evaluation. Multilinguality, Multimodality, and Interaction*. vol. 8685, E. Kanoulas, M. Lupu, P. Clough, M. Sanderson, M. Hall, A. Hanbury, *et al.*, Eds., ed: Springer International Publishing, 2014, pp. 116-127.

[14] R. Roller and M. Stevenson, "Applying UMLS for Distantly Supervised Relation Detection," in *Proceedings of the 5th International Workshop on Health Text Mining and Information Analysis (Louhi)*, 2014, pp. 80-84.

[15] P. Thomas, T. Bobic, M. Hofmann-Apitius, U. Leser, and R. Klinger, "Weakly Labeled Corpora as Silver Standard for Drug-Drug and Protein-Protein Interaction," *Third Workshop on Building and Evaluating Resources for Biomedical Text Mining Workshop Programme*, p. 63, 2012.

[16] M. Mintz, S. Bills, R. Snow, and D. Jurafsky, "Distant supervision for relation extraction without labeled data," presented at the Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 2 - Volume 2, Suntec, Singapore, 2009.

[17] S. Riedel, L. Yao, and A. McCallum, "Modeling relations and their mentions without labeled text," presented at the Proceedings of the 2010 European conference on Machine learning and knowledge discovery in databases: Part III, Barcelona, Spain, 2010.

[18] R. Hoffmann, C. Zhang, and D. S. Weld, "Learning 5000 relational extractors," presented at the Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics, Uppsala, Sweden, 2010.

[19] B. Rosario and M. A. Hearst, "Classifying semantic relations in bioscience texts," presented at the Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics, Barcelona, Spain, 2004.

[20] C. D. Manning, M. Surdeanu, J. Bauer, J. Finkel, S. J. Bethard, and D. McClosky, "The Stanford CoreNLP Natural Language Processing Toolkit," in *ACL Demonstrations*, 2014.

[21] Q. V. Le and T. Mikolov, "Distributed Representations of Sentences and Documents," *CoRR*, vol. abs/1405.4053, / 2014.

[22] A. M. Dai, C. Olah, and Q. V. Le, "Document embedding with paragraph vectors," *arXiv preprint arXiv:1507.07998*, 2015.

[23] R. Rehurek and P. Sojka, "Software framework for topic modelling with large corpora," in THE LREC 2010 WORKSHOP ON NEW CHALLENGES FOR NLP FRAMEWORKS, 2010, pp. 45--50.

[24] F. Bergenti, G. Caire, and D. Gotta, "Agents on the move: JADE for Android devices," in *Procs. Workshop From Objects to Agents*, 2014.

[25] J. P. Müller and K. Fischer, "Application Impact of Multi-agent Systems and Technologies: A Survey," in *Agent-Oriented Software Engineering: Reflections on Architectures, Methodologies, Languages, and Frameworks*, O. Shehory and A. Sturm, Eds., ed Berlin, Heidelberg: Springer Berlin Heidelberg, 2014, pp. 27-53.

[26] À. Bravo, J. Piñero, N. Queralt-Rosinach, M. Rautschka, and L. I. Furlong, "Extraction of relations between genes and diseases from text and large-scale data analysis: implications for translational research," *BMC Bioinformatics*, vol. 16, pp. 1-17, 2015

[27] W. Wu, H. Li, H. Wang, and K. Q. Zhu, "Semantic Bootstrapping: A Theoretical Perspective," *IEEE Transactions on Knowledge and Data Engineering*, vol. PP, pp.2016 ,1-1 .

Characterizations of Flexible Wearable Antenna based on Rubber Substrate

Saadat Hanif Dar, Jameel Ahmed
Department of Electrical Engineering
RIPHAH International University
Islamabad, Pakistan

Muhammad Raees
Department of Software Engineering
Mirpur University of Science and Technology (MUST)
Mirpur AJK, Pakistan

***Abstract*—Modern ages have observed excessive attention from both scientific and academic communities in the field of flexible electronic based systems. Most progressive flexible electronic systems require incorporating the flexible rubber substrate antenna operating in explicit bands to offer wireless connectivity which is extremely required by today's network concerned society. This paper characterizes flexible antenna performance under the environments developed by natural rubber as the substrate. Flexible antenna grounded on rubber substrate was simulated using CST microwave studio with diverse permittivity and loss tangent. In our work, prototype antennas were built using natural rubber with different carbon filler substances. This paper reveals advanced flexible substrate effects on antenna quality factor (Q) and its consequences on bandwidth and gain. Such antennas under bending washing environment were also found to perform better than existing designs, showing less change in their gain, frequency shift and impedance mismatch.**

Keywords—wearable antenna; antenna characterization; antennas

I. INTRODUCTION

Wearable and elastic wireless communication systems are gaining exceptional popularity due to their thoughtful prospective in daily life particularly in health care monitoring systems.Economic co-operation and Development (OCED) and BRIICS (Brazil, Russia, Indonesia, India and South Africa) are spending 6 % of their GDP to address long term health care issues and are anticipated to escalate 14 % in the next 50 years [1].Chronic disease identification and treatment through Remote patient monitoring (RPM) systems are the primary reasons of this increase.

Modern Flexible Electronics such as flexible mobile phones, electronics books and roll able keyboards are fitted out with an antenna to provide wireless connectivity .The competence of such systems depends on the characteristics of the integrated antenna near human body.Various design parameters of flexible wearable antenna is an important area that required to be evaluated as the different steps involved in its designing are relatively different from the rigid substrate oriented antennas [2].The broad design procedure of flexible and wearable antennas is depicted in figure.1

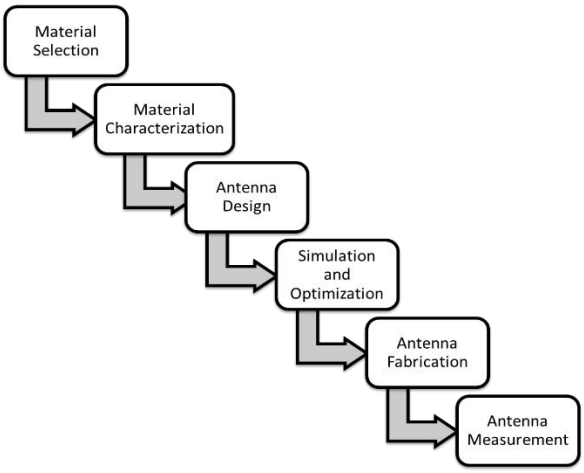


Fig. 1. Design procedure of flexible wearable antenna (Reproduced from [2])

The process starts with the selection and dielectric characteristics (electric conductivity, loss tangent, relative permittivity) of the conductive and substrate material.In the next step antenna geometry, ground plane, radiating element and feeding structure is defined. Finally, the critical parameters (durability, humidity, bending effects) are performed near human body.

Bendable wearable materials like conductive fibers, liquid metallic blends, and polymer in paper [3-6] are widely used in the existing flexible electronic devices. However, additional requirements are enforced in wearable applications. Therefore, these wearable applications require small size, light weight and low profile antennas, which must present stable electrical properties, low power consumption, reasonable impedance match and desirable radiation. Natural rubber is an attractive alternate flexible material which is biocompatible and offers high conductivity, low lost, ease to manufacture and the most important it is water/weather resistant and environment friendly.

Natural rubber is a very common material, however, its application in flexible electronics is very limited, and to our knowledge, flexible wearable antenna presented in [7] is the

first demonstration made so for to develop flexible antenna with rubber material. Rubber based substrates are insulators in nature since the atoms in rubber chain are covalent bonded. Conductive fillers such as carbon fibers or metal oxides are introduced to form the conductive paths in rubber [6,7].Carbon packings imparts a significant effect on the microwave characterization of natural rubber and hence increase in Tanδ and electrical conductivity ,as reported in [8].

In this paper, Natural rubber is characterized to design flexible wearable antenna for on-body communications.We conclude our study by reporting characterization of actual antenna prototype, including details of fabrication processes, dielectric properties of the substrate and the consequence of filler contents on antenna quality factor (Q).Further to this, different challenging factors for flexible antennas like bending, wrinkling, Wash ability and environmental factors (humidity and thermal effects) that effects on antenna efficiency and gain in addition to the return loss, radiation pattern are also investigated.

II. Selection Criteria for Substrate

Body area networks (BAN's) require effective wireless connectivity to integrate wearable textile antennas in flexible electronic systems. Rubber substrate is a flexible material and its thickness might change with low pressure.

A. Permittivity

The permittivity, ε is a multifaceted parameter and its relative permittivity is formulated as

$$\varepsilon_r : \varepsilon = \varepsilon_0 \ \varepsilon_r = \varepsilon_0(\varepsilon'_r - j\varepsilon''_r)$$

whereas ε_0 represents permittivity of vacuum, which is $8.854 \times 10{-}12$ F/m [9]. Dielectric constant of the various flexible substrate lies in the range of $2.2 < \varepsilon_r < 12$ [10].

The lower dielectric constant declines the surface wave losses which are tied to guided wave broadcast within substrate. Therefore; lower dielectric constant raises the impedance bandwidth of the antenna with adequate competence and high gain [11].

Fig.2 demonstrates the deviation of ε with frequency for diverse filler contents in rubber substrate. Significant change in ε'_r is observed by changing filler contents from 0% to 60%. Relative permittivity is enhanced when the filler contents are increased.EM signals can move through the material easily and henceforward this will result to lower the permittivity. As obvious from the figure, by increasing carbon filler in natural rubber structure of the substantial material becomes more dense and porous.

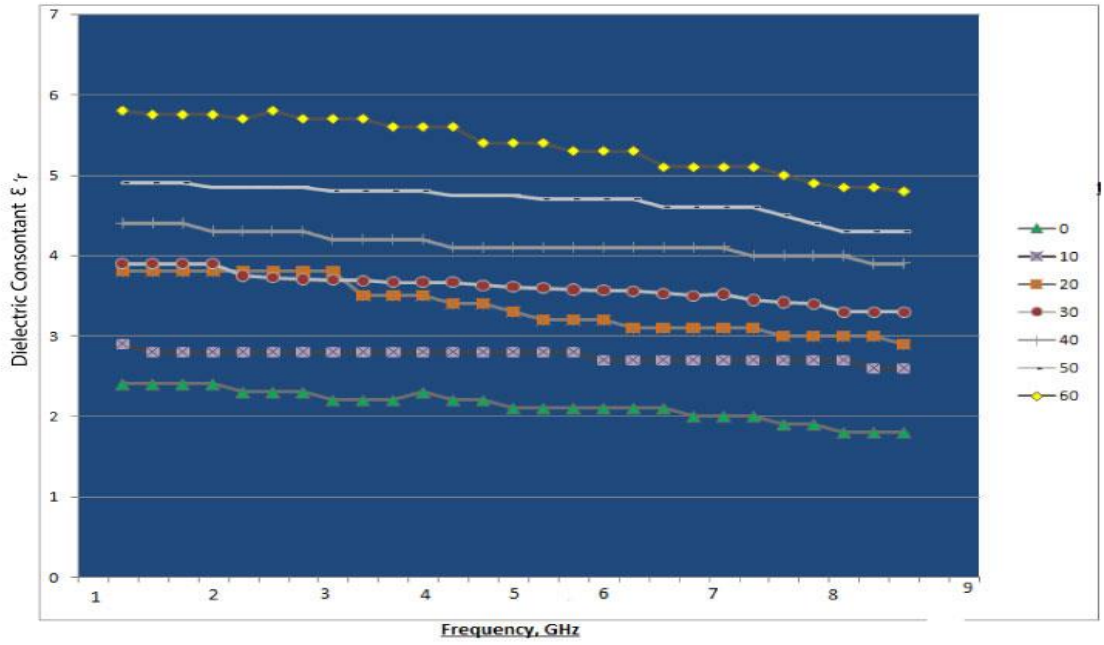


Fig. 2. Relative Permittivity εr relation with carbon filler components

B. Loss Tangent

Loss Tangent (tanδ) is also identified as dissipation factor. It describes the amount of power turned into heat in the substantial material. Loss tangent in the following relation is defined as the ratio of the imaginary part to real part of the relative permittivity.

$$Tan\delta = \varepsilon'' / \varepsilon_r' \quad (1)$$

The higher values of loss tangent results in additional losses in the dielectric substrate and higher losses outcomes in reduced radiation efficiency. The higher losses in tangent values lead to the more losses in dielectric substrate [12] and as a result antenna efficiency reduces.

Figure 3 plots deviation of relative permittivity against the frequency.As perceived from the figure, by increasing the rubber filler contents significant increase in tanδ is observed.

The statement that with the addition and increment of rubber filler contents material gets loss and this leads to an increase in electrical conductivity, reported in [12].Since the carbon contents impacts electrical conductivity and as a result the dielectric properties of the rubber and hence effects on bandwidth, return loss and quality factor of antenna.

C. Thickness of the Dielectric Fabrics

The bandwidth and competence of a natural rubber based flexible antenna is principally decided by the substantial substrate dielectric constant and its width. The width h of substrate normally lies in the range of $0.003\ \lambda \leqslant h \leqslant 0.005\ \lambda$ whereas λ represents its wavelength. For a fixed comparative permittivity, the substrate thickness may be selected to maximize the bandwidth of the flexible antenna. Though, this value may not improve the antenna efficiency.

The effect of the thickness on the bandwidth (BW) of the antenna could be described by Equation (2), where Q is the antenna quality factor.

$$BW \sim 1/Q \tag{2}$$

The quality factor (Q) is prejudiced by the space wave losses, surface wave and dielectric losses. The selection of the thickness of the dielectric substantial material is a negotiation between competence and bandwidth of the antenna.

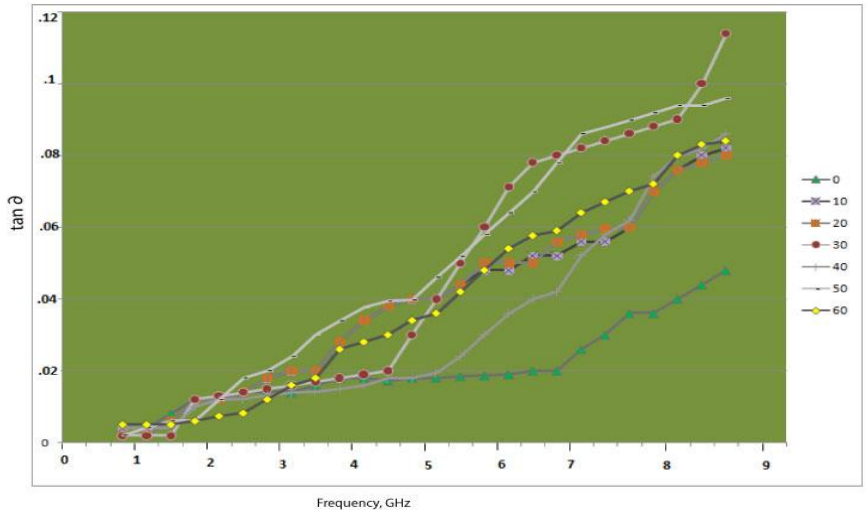


Fig. 3. Tangent loss (Tanδ) relation with frequency on different carbon compositions

III. Antenna Design Criteria

Micro strip patch antenna design is used to present natural rubber based flexible wearable antenna.Several types of antennas like micro strip patch antennas were designed by using rubber in CST Microwave Studio Software. Wire antennas in stand-alone approach experiences shift in frequency due to variation of wavelength, which depends on distance from human body. However, rubber based flexible antenna experiences less effects in on-body communications. Its major effect is related with the location of antenna on the body and type of antenna being used. Natural rubber based patch antenna due to full ground plan reduces back radiations while placing near human body [13]. It is exciting that Flexible antennas designed with natural rubber are suitable candidate for wearable antennas in Body Area Networks (BAN's).

The geometry of Flexible Antenna shown in figure.4 aims to investigate wearable patch antenna with different height of substrate.Each antenna had the same path size however; its relative permittivity is varying. The antenna feature a simple copper patch which act a radiator, fed by 50Ω inset feed line. Antenna dimensions are calculated by using the transmission line model based on substantial substrate permittivity and loss tangent of 3.2 and 0.01.

To design a patch rubber based antenna approximate value of dielectric is taken into account. The dielectric constant value of the natural rubber substare may be calculated by simply measuring the resonant frequency of small area radiator. Rubber patched antennas are designed by calculating its various dimensions. The patch width (w) imparts a slight effect on the resonant frequency (fr), and it is designed by using the following formula [14].

$$w = \frac{c}{2fr}\sqrt{\frac{2}{(\varepsilon r+1)}} \tag{3}$$

Width (W) and length (L) of the patch are the major ingredients which characterizes the antenna design. Width of the radiating patch imparts slight effect on the radiation pattern shapes, but has major effect on the input impedance and operating bands.

Radiation power of the antenna increases as the width of the radiator in increased. As a result ,it expands bandwidth and increases efficiency. It is supposed that ratio of width to length lies in the range of $1 < W/L < 2$.

Width (W) and length (L) parametric values satisfy the above conditions in the following equations:

$$L = \frac{c}{2fr\sqrt{\varepsilon re}} - 2\Delta L \tag{4}$$

Where, εre is the effective dielectric constant and is calculated by using the following relation

$$\varepsilon re = (\varepsilon r+1)/2+(\varepsilon r-1)/2(1+12\ h/w)1/2 \quad (5)$$

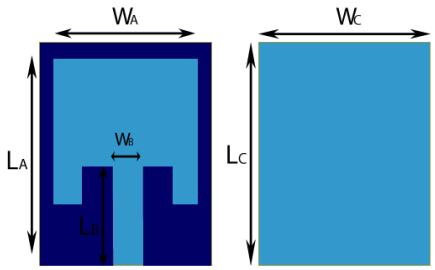


Fig. 4. Antenna Design (a) front view and (b) back view

IV. Results and Discussions

To Study the effect of rubber substrate, flexible wearable antenna simulations were conducted by using commercially existing CST studio to disclose several dielectric properties, and further to calculate their impacts on the antenna performance in terms of quality factor (Q).

Figure 5 and Figure 6 shows the simulated and measured rejoinders of the antenna correspondingly. The following assumptions can be made from the found results.

1) The band-width increases with the addition of carbon contents in the rubber substrate.As the carbon contents increases, Tanδ increases and the rubber substrate gets loss and hence the Q decreases.

2) The return loss degraded with the accumulation in rubber contents. As, the rubber contents increases, the substrate resistivity and permittivity changes, leading to variation in impedance match.

3) The resonant factor is also affected by carbon contents.The shift in frequency is enlightened by the act as the permittivity changes, the wavelength variations as well, giving increase to frequency.

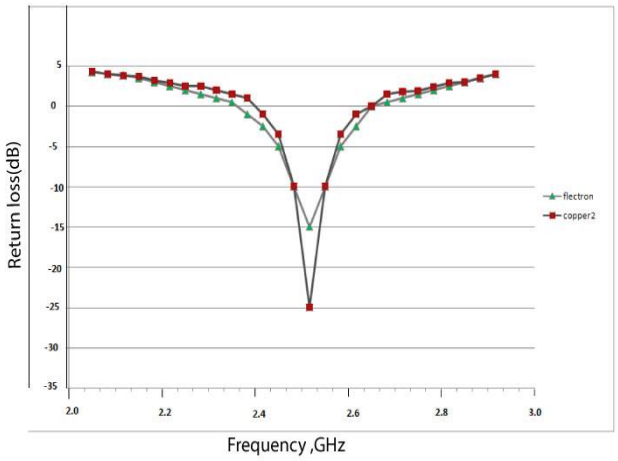


Fig. 5. Simulated return loss of the proposed antenna with diverse contents of rubber

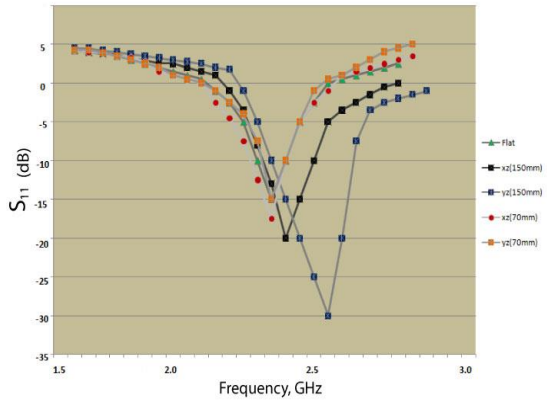


Fig. 6. Simulated return loss of optimized antenna

The prototype antenna is analyzed based on return loss and gain. The bandwidth for the antenna is increased as Tanδ of the rubber substrate were increased. Characteristically, the total Q is given by the following equation.

$$1/Q = 1/Q_{rad} + 1/Q_d + 1/Q_c + 1/Q_{sw} \quad (6)$$

Where, Q_{rad}, Q_c, Q_d and Q_{sw} are the performance parameters of flexible antenna .The Q of the dielectric (Q_d) is inversely proportional to Tanδ of dielectric concluded in equation 7.

$$Q_d = 1/ Tan\delta \quad (7)$$

Therefore, the rise in rubber contents causes the increase of electrical conductivity of substrate and as a result increases in Tanδ. Finally, the bandwidth is increased because bandwidth (BW) is inversely proportional to Voltage-standing-wave-ratio(VSWR),shown in equation 8.

$$BW=(VSWR-1)/(Q \surd VSWR) \quad (8)$$

V. Opportunitsic Challenges of Wearable Antenna

A. Bending and crumpling effects

When wearable flexible antennas are operated on human body, bending, crumpling and sometimes, twisting actions are inevitable.For practical applications, the changes impart adverse effects on the antenna performance.

Detailed technique on the flexible antenna tests reported in [15] is briefed as follows:

1) Robustness and strength tests are executed by repeated trials of the fabrication antenna under bending, crumpling, and twisting to guarantee there are no wrinkles which might reduce the antenna's performance.

2) Resonant frequency and return loss are characterized under twisting conditions since they are inclined to weaken due to impedance discrepancy and capacitive coupling. Bending tests are conducted by confirming the antenna under test on foam cylinders of different radii to parodist diverse bending degrees. The impedance matching declines with a shift in operating frequency as the flexible antenna under goes bending and crumpling circumstances.

3) On the other hand, wrinkling formers are usually engaged to measure the antenna performance under bending and crumpling conditions. The crumpling results in by varying in different directions. The impedance matching declines with a shift in operating frequency as the flexible antenna under goes bending and crumpling circumstances.

The radiation efficiency varies consequently and therefore radiation patterns are distorted. The volume of distortion signal is dependent on the magnitude of bending and crumpling, as reported in [16].

B. Extensive Tests for Flexible Wearable Antenna

Supplementary measurements are needed when flexible antenna is operated in a specific situation. Here, two extensively conducted tests are introduced:

1) Washing effects (Washability)

Wearable bendable textile-based antennas are typically uncovered to dust, dirt and excretion, which might compromise their performance. Furthermore, wearable Textile antennas that are merged within clothing are subject to be soaked with water and or wash away. The performance of antenna is obligatory to be consistent after it is washed. To measure the antenna performance after going through wet conditions, its Wash ability test is reported in [17].Apparently the antenna performance in such conditions depends on the conductive and substrate substantial choice.

2) Atmospheric effects (Moisture and Temperature tests)

Wearable flexible Martials are sensitive to the environmental variables, such as moisture and temperature. As in more or less suitcases wireless systems are required to operate under tough environmental conditions. Performance constancy tests are indispensable.

When an antenna is operated in moist and hot weather atmosphere, the impact of relative humidity and temperature should be examined.

In [18], the effect of moisture and dampness on the reflection coefficient are explored by varying humidity level from 10% to 90 %.Apparently, with increased relative moisture, the permittivity and the loss tangent are both amplified.

VI. Conclusion

This Paper has presented a thorough study on flexible wearable antenna built from a new substantial substrate (rubber).Antenna prototype was presented and its performance near human body is investigated.

The effect of rubber contents on antenna quality factor and bandwidth were pragmatic. With enhanced processing techniques, it should be possible, in theory, to produce rubber with better dielectric properties, and therefore antennas with better performance should be appreciated. It is observed that bandwidth and return loss are improved by adding rubber contents .The antenna efficiency and hence its gain decreased, but this is considered reasonable since natural rubber is quite loss in nature. With improved processing techniques, it should be possible, in theory, to produce rubber with better dielectric properties, and as a result antennas with better performance should be realized.

However, it is pertinent to mention that the antenna performed better when they were subjected to bending and washing conditions than antennas with other substrate. In summary, we have proved the viability of using a new natural material in designing bendable antenna that can tolerate under washing and other environmental conditions like humidity. We have established that antenna performance can be well-ordered by adding rubber contents.

References

[1] de la Maisonneuve, C., Martins, J.O., "Public spending on health and long-term care: a new set of projections,"OECD Economic Policy Papers, no. 6. (2013) http://www.oecd.org/economy/public-spending-on-health-and-long-term-care.htm .

[2] H.Khaleel, Innovation in Wearable and Flexible Antennas:, Ser.Wit Transaction on State-Of-the-art in Science and Engineering, WIT press, 2015.

[3] Liyakath, R. A., A. Takshi, and G. Mumcu, "Multilayer stretchable conductors on polymer substrates for conformal and reconfigurable antennas," IEEE Antennas and Wireless Propag. Lett., Vol. 12, 603–606, 2013.

[4] Hayes, G. J., A. Qusba, M. D. Dickey, and G. Lazzi, "Flexible liquid metal alloy (EGaIn) microstrip patch antenna," IEEE Trans. Antennas Propag., Vol. 60, No. 5, 2151–2156, May 2012.

[5] Hazra, R., C. K. Ghosh, and S. K. Parui, "Effect of different semi conductive substrate materials on a P-shaped wearable antenna," Int. J. of Adv. Res. in Comp. and Comm. Eng. (IJARCCE), Vol. 2, No. 8, 3071–3074, 2013.

[6] Xi, J., H. Zhu, and T. T. Ye, "Exploration of printing-friendly RFID antenna designs on paper substrates," IEEE Int. Conf. on RFID, 38–44, April 2011.

[7] Zaiki Awang, Nur A.M.Affendi and Nur M.Razali, "Flexible Antenna Based on Natural Rubber,"Progress In Electromagnetic Research C, Vol.61,75-90,2016.

[8] Olivera,F.A.,N.Alves,J.A.Giacometti,C.J.L Constantino, L.H.C. Mattoso, A.M.O.Balan,and A.E job, "Study of the thermomechanical and electrical properties of conducting composites containing natural rubber and carbon black,"J.Appl.Polym.Sci.,Vol.106,No.2,1001-1006,2007.

[9] Rais,N.H.M.Soh,P.J.Maliek,F.Ahmad.S.Hashim,N.B.M.Hall,P.S "A review of Wearable Antenna",Antenna & Propagation Conference,2009,pp 225-228.

[10] C.A.Blannis, "Antenna Theory :Analysis and Design",3rd ed.,,Wiley,2005,pp.770.

[11] Baker-Jarvis ; Janezic, M.D.; DeGroot, D.C. High-Frequency Dielectric Measurements. IEEE Trans. Instrum. Meas. 2010, 13, 24–31.

[12] B.Gupta,S.Sankarlinngam,S.Dhar , "In proceedings of Mediterranean Microwave Symposium (MMS), Turckey,2010,pp.251-267.

[13] Affendi, N. A.M., N. A. L. Alias, Z. Awang, M. T. Ali, and A. Samsuri, "Microwave non-destructing testing of rubber at X-band," 2013 IEEE Int. RF Microw. Conf., 333–337, Penang, December 2013.

[14] Jaime G.Santas, Akram Alomaniny and Yang Hao, "Textile Antenna for on-body communications: Techniques and properties",EuCAP,2007

[15] Khaleel,H.R.,Al-Rizzo,H.&Rucker,D.,Compact polyimide based antennas for flexible displays.IEEE journal of Display Technology,8(2),pp.91-97,2012

[16] Bai,Q.&Langley,R.Crumpling of PIFA textile antenna.IEEE Transaction on Antennas and Propagation,60(1),pp.63-70,2012

[17] Scarpello,M.,Kazani,I.Hertleer,C.,Rogier,H.&Ginste,D.,Stability and efficiency of screen-printed wearable and washable antennas.IEEE Antennas and wireless Propagation Letters,11,pp.838-841,2012

[18] Hertleer,C.,Van Laere,A.,Rogier,H.&Van Langenhove,L.,Influence of relative humidity on textile antenna performance.Textile Res.J.,80(2),pp.177-183,2010

E-Commerce Adoption at Customer Level in Jordan: an Empirical Study of Philadelphia General Supplies

Mohammed Al Masarweh
Software Engineering Department
Al-Balqa' Applied University
Salt, Jordan

Laila Al-Qaisi
Computer Science WISE
Amman, Jordan

Sultan Al-Masaeed
E-Business Department
Al Ahliyya Amman University
Amman, Jordan

Ziad Hunaiti
ECE department
Brunel university London
London, UK

***Abstract*—E-commerce in developing countries has been studied by numerous researchers during the last decade and a number of common and culturally specific challenges have been identified.. This study considers Jordan as a case study of a developing country where E-commerce is still in its infancy. Therefore, this research work comes as a complement to previous research and an opportunity to refine E-commerce adaptation research. This research was conducted by survey distributed randomly across branches of Philadelphia General Supplies (PGS), a small and medium enterprise (SME). The key findings in this research indicated that Jordanian society is moving towards online shopping at very low rates of adoption, due to barriers including weak infrastructure throughout the country except in the capital, societal trends and culture and educational and computer literacy. This means that E-commerce in Jordan still remains an under-developed industry.**

***Keywords*—Information systems; E-commerce; E-commerce Adoption; E-commerce in Jordan; Jordan**

I. Introduction

Information Technology (IT) has been significantly developing in recent decades, raising numerous opportunities and challenges in the business environment in various aspects. In a consumer context, the synergy of the internet and business operations gives more flexibility to users to identify their needs as well as respond to market developments, such as feedback and reviews about items customers have bought and their experiences [1].

Since the 2000s the popularization of personal computing and the revolution in internet communications has enabled a massive adoption of E-commerce throughout the world, correspondingly making it a major concern or researchers exploring the consumer and business implications of associated developments. While E-commerce is now firmly embedded in the economic life of most developed countries, its progress has been uneven and more complex in developing countries, thus most studies pertaining to the latter have focused on adoption factors, with a general view to exploring how E-commerce can be enhanced to promote socio-economic development.[2]

Within the global economy, essentially all financial transactions are now conducted online, and a virtual market has arisen in commercial technological innovation that supports a new channel of client-supplier transaction, which has been widely used in business-to-business transactions to greatly enhance operational cost and time efficiency; however, there is a disconnect between the significant streamlining of business-to-business operations (such as those involved in supply chain management) and business-to-customer transactions in the electronic environment, with salient differences in interactions and customer patterns.[3]

Most organizations tend to abandon traditional business formats as much as feasible where E-commerce alternatives can be adopted, but such restructuring requires coherent and appropriate strategies. However, to remain competitive in the global marketplace (and indeed in basic national economic contexts) companies of all kinds have been compelled to adopt E-commerce applications [4].

Corresponding to global trend of moving organizations into E-business, many organizations in the Hashemite Kingdom of Jordan (HKJ) had already moved, while others are studying their steps towards moving [2].

Harnessing the power of the Internet could help SMEs to address critical issues, such as: increased global competition and consumer demand for quality, rapidly changing market environment, growing need for flexibility and immediate access to business information [5].

The main concern in this study is E-commerce adoption for small and medium enterprises (SMEs) in the HKJ, investigated by an empirical study of one of those SMEs, namely Philadelphia for General Supplies.

II. Literature Review

A. Electronic Commerce

E-commerce generally refers to the integration of IT with various business activities within an organization. This plays critical role in spreading the organization's services and products beyond its local market, as well as coordinating

internal operations and supply chain management as a whole. As a result, it helps in gaining more customer interest and other various benefits affecting targets positively [6].

Many researchers have defined E-commerce from various perspectives. Khoshnampour and Nosrati [7] gave a simple definition about selling and buying products and services through the internet or other types of networks, while [6] indicated that it is about using the internet to accomplish business transactions locally or internationally. The use of IT to conduct business transactions among buyers, sellers and other partners was the crux of the definition presented by [8], while [9] gave a more specific and subjective definition of using Web 2.0 applications to conduct undergoing commerce activities in order to improve customer participation, which results in improved satisfaction and greater economic value.

Turban et al. [10] defined E-commerce as a business model in which transactions are conducted through electronic networks (i.e. the internet), including the processes of buying and selling products, services and information. Nielson et al. [11] stated that a company can be seen as a true E-commerce venture when its major revenue is generated via the internet; the online environment is connected to all of its major processes; business operations can be undertaken 24/7; and they are always open to satisfy global customers. Secondary features include organizational structure being less centralized and less hierarchical than in traditional business models. In having these traits, a company can react to rapid changes in the digital world very quickly, whereby flexibility and the ability to execute changes are highly required.

B. E-commerce adoption within an organization

The realm of business has been profoundly revolutionized by the global spread of electronic communication technologies, particularly the internet, and their popularization and widespread use since the late 1990s. Prior to this, only large multinational corporations could afford to invest in hardware, software and skilled personnel to achieve efficiencies from investment in IT, but the increasing affordability of personal and industrial computer systems has driven the adoption of IT and E-commerce in society at large, with organized E-commerce systems being essential to SMEs and even small companies in the modern market [12].

The work of Lippert and Govindarajulu [13] showed that successful adoption of E-commerce by any organization is based on two main factors: technology characteristics and customer understanding of the system. Nevertheless, adoption decisions within an organization are conditioned by multiple complex and often interconnected wider stakeholders in the surrounding environment, including customers, suppliers, partners, competitors and governmental regulations. This means that top managers have to study their organizational environment's current state and decide whether it is possible and economically viable to complete the adoption process, on the basis of feasibility and cost-benefit analysis [13].

Numerous different procedures are necessary for different E-commerce technology adoptions, such as websites enabling online business transactions, which require a payment platform as well as catalogue and ordering systems linked to inventory and ordering functions. E-commerce adoption must therefore be taken seriously, with full acceptance of its complexity and potentially high investment costs from the outset. Empirical analyses have generally explored the process of adoption in terms of sequential stages, from initial awareness of the potential innovation to full integration and deployment of E-commerce functions in organizational practices [14] and [15].

Furthermore, [16] mentioned that organizations should be aware of the necessity for a dedicated online marketing strategy in addition (or in place of) traditional marketing methods, to utilize the advantages of E-commerce to reach wider or (conversely) more targeted audiences more expediently. Traditional mass advertising, while remarkable in its day, pales in comparison to the advertising capabilities of technologies and platforms enabled by web 2.0 applications, such as social networking websites, customer reviews and suggestions, and the ability of personalizing advertisements through such interactive websites. E-commerce enables properly organized companies to seamlessly combine the functions of market research, advertising and supply chain management with online purchasing activity, with maximum cost efficiency and the most effective deployment of resources.

C. E-commerce adoption challenges

Despite the manifest advantages of E-commerce adoption, it cannot simply be purchased and implemented wholesale; for effective deployment, organizations must devise an adoption strategy tailored to their own needs, objectives and surrounding context. One of the main challenges is to define an effective E-commerce model and strategy [17] and [18]; this must be done with reference to three main components: community, content and commerce , relative to the organization's industry, services and products. Chat rooms, message boards, email lists and social networking websites have been used to gain community (i.e. a large number of interested and motivated potential customers). Content includes news, descriptions or any related information that can attract potential customers, as well as customer-generated content such as product reviews, which disseminates information among their community. Commerce is represented by consumer purchase decisions for physical products or services online [17][18].

Another challenge that might appear during this migration, especially when organizations attempt to move into business-to-customer (B2C) E- commerce, is to integrate a web-based front-end ordering system with the back-end computer-based systems relating to inventory control and production planning. The web based front-end customers interface should be tightly integrated with back-end inventory computer-based systems in order to ensure the availability of requested products, and to alert the purchasing department to production planning needs, as shown in Figure 1 [18]. Other challenges that might appear are security and trust; internet experience; language; legal issues; and technology acceptance [19]

D. E-commerce adoption in Jordan

A few studies have focused on E-commerce adoption in Jordan; their main findings are presented in this section. The current status of E-commerce in Jordan represents a vital factor affecting this research. Al-debei [20] found that Jordan has various adequate E-commerce prerequisites, such

Identify applicable sponsor/s here. If no sponsors, delete this text box (*sponsors*).

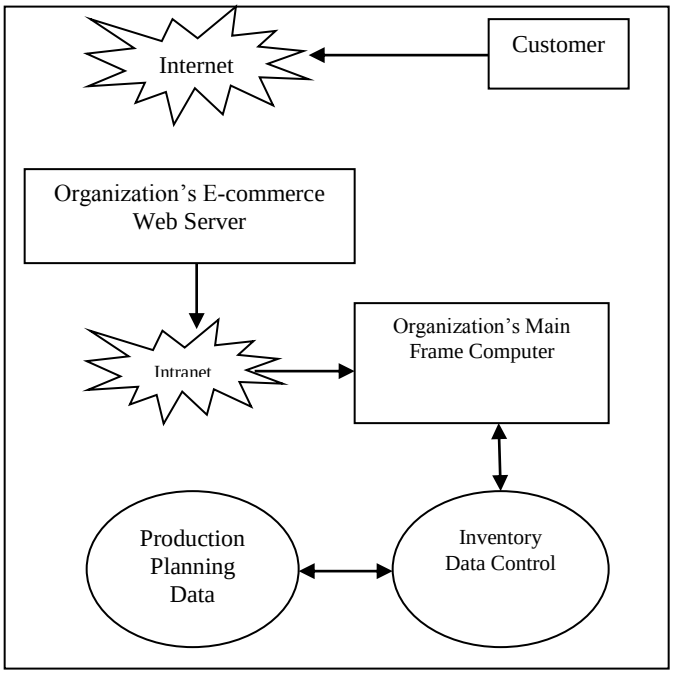


Fig. 1. Integrating web based front-end ordering system with the back-end computer based systems

as technology and telecommunication infrastructure, organizations readiness and support, institutional and governmental support and community culture. Moreover, there are sufficient IT vendors who can provide Jordanian companies with required hardware and software for their move towards E-commerce. However, the study identified that Jordanian community culture is not germane to E-commerce adoption, and cultural beliefs and practices constitute a major impediment to this major change in the local business market. This was cited as the main reason why relatively few organizations (he specified less than 20) have developed E-commerce in Jordan [20].

Alamro and Tarawneh [21] presented a model of the main factors in E-commerce adoption in Jordan based on three dimensions: the External Environment, the Organizational Context and the Technological Context, as shown in Figure 2.

A 2015 study [22] concluded that the effect of the adoption of e-commerce systems by SMEs in Jordan is affected particularly by the following factors: readiness, strategy, managers' perceptions and external pressure by trading partners.

A survey questionnaire was used to evaluate this model, finding that the external environment context (e.g. customer demand and quality of IT consultation services) are the main factors that affect E-commerce adoption in Jordan. In an organizational context, it has been found that employees' ICT knowledge and attitudes affect E-commerce adoption, making them either facilitators or inhibitors. In the technological context, a lack of trust in banks' support for electronic transactions, the threat of disintermediation and fear of identifying theft were the main inhibitors.

In conclusion, many factors militate against organizations' move towards E-commerce in Jordan, mainly attributable to cultural factors and a lack of trust in the security of E-payment systems.

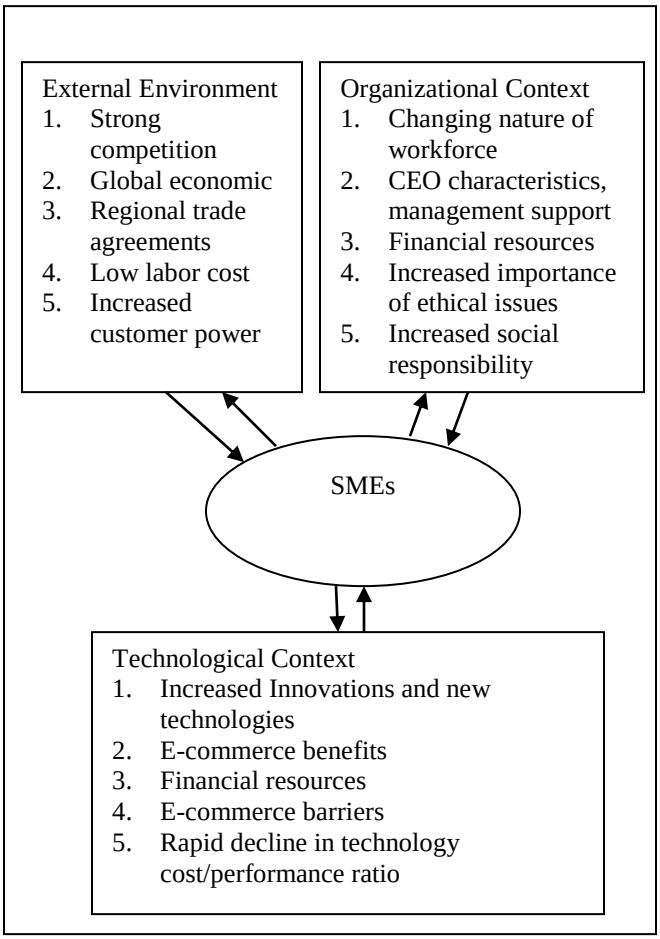


Fig. 2. A model of E-commerce adoption in SMEs

III. Research Methodology

A. Approach

The preceding literature review about the current status of E-commerce in Jordan has given a general idea about the main concern of this study, and indications of where to focus to explore the main factors affecting E-commerce adoption for Jordanian companies.

A set of interviews with PGS management directed the researcher to conduct a questionnaire to be answered by their customers in order to examine E-commerce adoption at a customer level, as their online shopping experience indicates the extent to which extent this major change will affect PGS's adoption of E-commerce in order to achieve intended goals. Hajli et al. [23] conducted a similar study entitled Examining E-commerce Adoption at Customer Level: The Impact of Social Commerce in Iran; his questionnaire was found to be very suitable and useful to be customized and used for measuring Jordanian customers' behavior towards online shopping.

B. Research design

This research was conducted based on the main factors identified from a similar study in a different developing country in the Middle East. Interviews were held with top management personnel to customise the main factors for inclusion in the survey administered to intended customers in order to achieve the main goal of the study.

Questionnaire variables were concluded from interviews with top management executives. Three different locations for PGS stores were selected to distribute the questionnaires to their existing customers. The survey was designed to focus mainly on customers' opinions about online shopping and to examine their own experiences in this regard. In order to study a wide range of customers' behavior, different cultures and different age ranges were targeted by the distribution of 200 questionnaires in three locations, namely Jab AlHussien, Down Town and Al-Wehdat branches of PGS. A period of one week was given for participants recruited from each branch to return their questionnaires. A total of 189 completed responses were received, a response rate of 94.5%. The results were then analyzed using SPSS. The results of statistical (quantitative) and thematic (qualitative) analysis are presented in the next chapter. Figure 3shows the research design for this study.

C. Questionnaire design

The questionnaire consists of three types of questions: multiple choice,30 Likert-type questions and finally two open-ended questions to address major challenges towards E-commerce adoption in Jordan from customers' perspectives.

The multiple choice questions focused on if and how customers use the internet to shop online. The Likert-type scale question addresses respondents' ability to search about requested products online, their trust of visited online shopping websites, familiarity with online shopping, familiarity with social networking websites and communities online, trusting others' reviews of certain products, the ability to provide the online vendor with required information to better serve customers' needs (i.e. market research) and respondents' opinions about learning how to use computers, browse products on the internet and shop online.

Furthermore, it addressed respondents' opinions of websites' flexibility, simplicity, ease to interaction and sociability, as well as the particular issue of their feelings concerning using credit cards to shop online.

D. Evaluation of results

The survey aimed to understand the attitudes of Jordanian customers towards E-commerce adoption and their ability to shop online. The results of this questionnaire helped to have a broad overview of PGS customers' experiences and perspectives towards online shopping as a rudimentary example of general Jordanian E-commerce customers (actual and potential). Moreover, measuring customers' ability to understand and deal with online websites, the internet, computers and Social Network Service (SNS) generally would indicate their willingness to deal with PGS's migration process towards E-commerce.

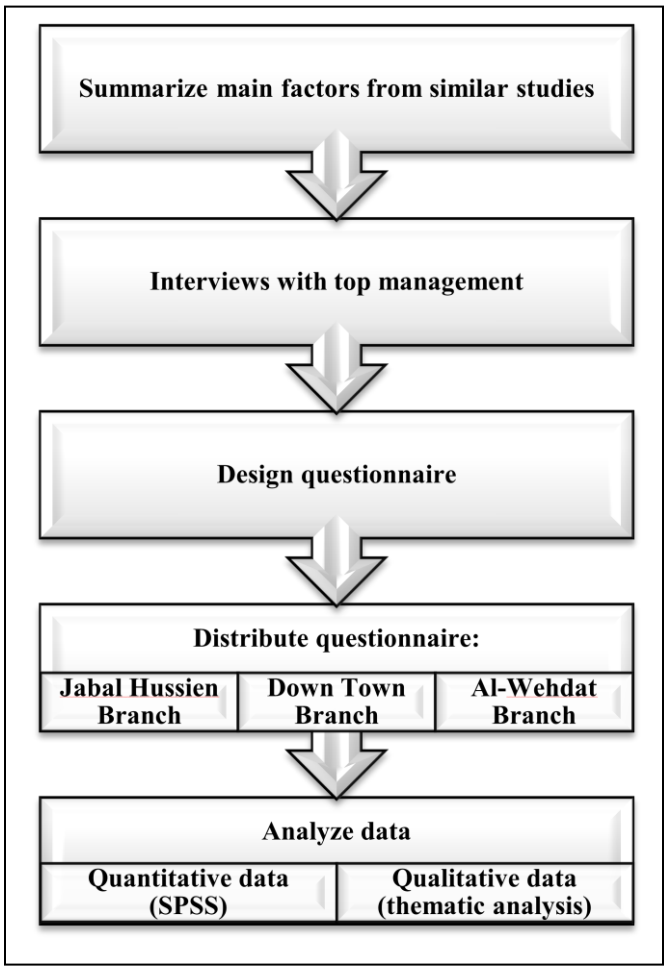


Fig. 3. Research design

IV. Results

As explained in the previous chapter, a survey questionnaire was administered been conducted to measure E-commerce adoption at the customer level in Jordanian society, considering customers of PGS as a case study of Jordanian E-commerce customers. This helped in understanding the current status of online shopping in Jordanian society and the extent to which Jordanian organizations would benefit from developing their E-commerce activities. Most of the survey items comprised Likert-type questions, an example of which is illustrated below in Table I.

TABLE I. Example of Likert Statement Used in this Research

Promises made by the website that has used for the last online shopping are likely to be reliable				
Strongly agree	***Agree***	***Neutral***	***Disagree***	***Strongly disagree***
1	2	3	4	5

A. Reliability

The Cronbach's (t-test) alpha test was used to ascertain the reliability of this research. The Cronbach's alpha value was 0.906, which is above the 0.70 threshold, indicating adequate internal consistency.

B. Participants description

Participants comprise a convenience sample of190 adults, all of whom (100%) are Jordanian citizens, comprising 126 females and 74 males, as illustrated below in Figure 4.The majority of participants were aged 20-25 years old, as illustrated in Figure 5, and the majority held a bachelor's degree and have no online shopping experience, as illustrated in Figures 6-7.

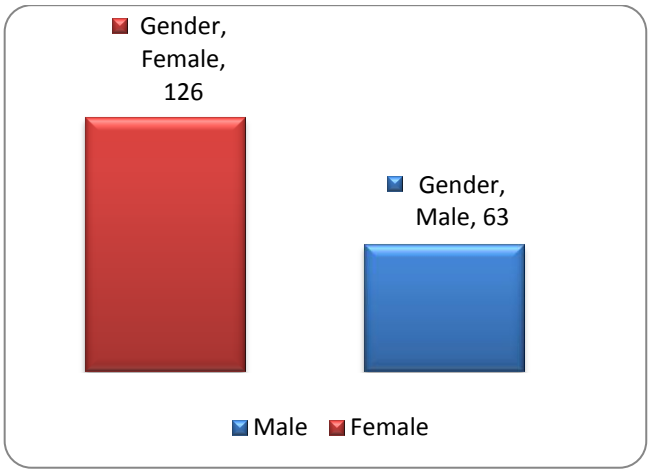


Fig. 4. Sample gender

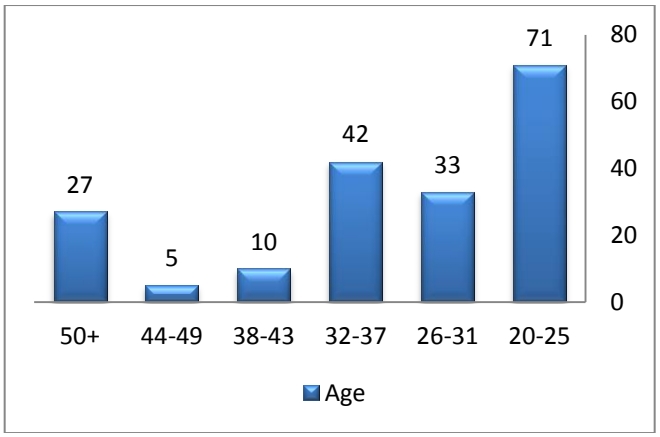


Fig. 5. Sample age (yrs)

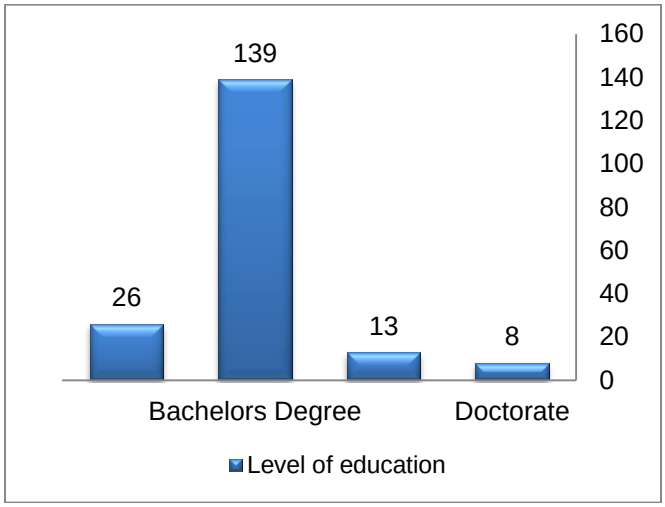


Fig. 6. Sample level of education

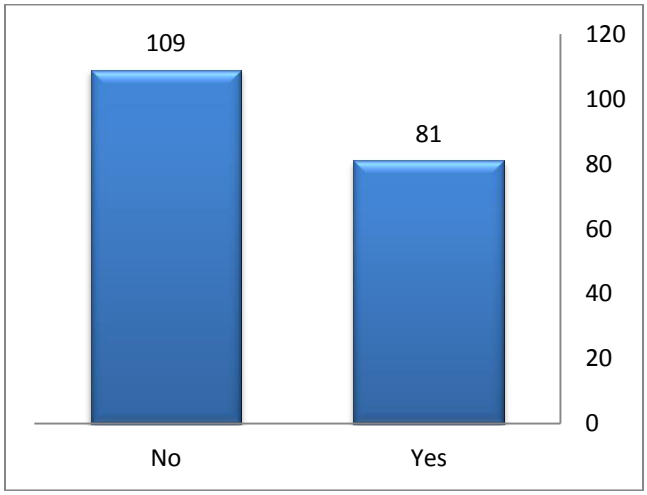


Fig. 7. Sample experience of online shopping

C. Ability to search about desired product online variable

Results indicate that Jordanians agree that searching and buying on the internet is useful, and they are familiar with internet searching for shopping purposes.

TABLE II. SECTION 1 – ABILITY TO SEARCH ABOUT DESIRED PRODUCT ONLINE

Strongly agree		Agree		Neutral		Disagree		Strongly disagree		Mean	SD
N	%	N	%	N	%	N	%	N	%		
1. Searching and buying on the internet is useful for me											
0	.0	6	7.4	13	16.0	24	29.6	38	46.9	4.16	0.96
2. Searching and buying on the internet makes my life easier											
2	2.5	2	2.5	15	18.5	36	44.4	26	32.1	4.01	0.92
3. I am familiar with searching for materials on the internet											
2	2.5	1	1.3	22	27.8	29	36.7	25	31.6	3.94	0.94
4. I am familiar with buying materials on the internet											
0	.0	8	9.9	21	25.9	37	45.7	15	18.5	3.73	0.88
Total ability to search about desired product online										3.96	0.63

Table Ⅱ and Figure 8 show that the total of means of the variable total ability to search about desired product online was (3.96), with standard deviation (SD) (0.63), which means this ability was low. Also, it can be noted that question 2 (Q2) "Searching and buying on the internet makes my life easier" had the highest mean (4.16, SD0.96), while Q4"I am familiar with buying materials on the internet" had the lowest mean (3.73, SD0.88) (Table Ⅱ).

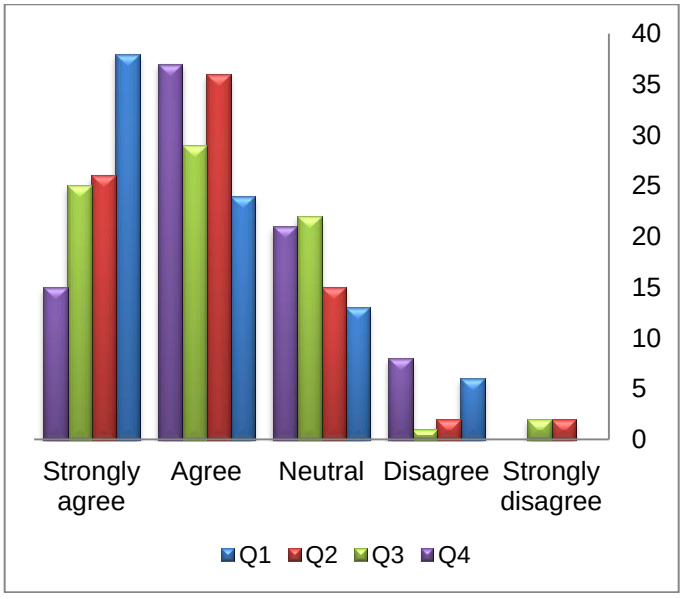


Fig. 8. Ability to search about desired product online

TABLE III. SECTION 2 – CUSTOMER'S TRUST IN USED ONLINE SHOPPING WEBSITES

Strongly agree		Agree		Neutral		Disagree		Strongly disagree		Mean	SD
N	%	N	%	N	%	N	%	N	%		
5. Promises made by the website that I used for my last online shopping are likely to be reliable											
2	2.5	4	5.0	26	32.5	34	42.5	14	17.5	3.68	0.91
6. I do not doubt the honesty of the website that I used for my last online shopping											
1	1.3	11	14.1	18	23.1	28	35.9	20	25.6	3.71	1.05
7. The websites on the internet enable me to search and buy materials faster											
0	.0	4	4.9	8	9.9	34	42.0	35	43.2	4.23	0.83
8. The websites increase my productivity in searching and purchasing products on the internet											
1	1.3	7	8.9	16	20.3	38	48.1	17	21.5	3.80	0.93
12. Based on my experience with the online vendor in the past, I know it is honest											
0	.0	5	6.2	18	22.2	39	48.1	19	23.5	3.89	0.84
13. Based on my experience with the online vendor in the past, I know they care about customers											
2	2.5	5	6.2	27	33.3	24	29.6	23	28.4	3.75	1.02
14. I am very likely to provide the online vendor with the information it needs to better serve my needs											
1	1.2	8	9.9	14	17.3	32	39.5	26	32.1	3.91	1.00
15. I usually use people ratings and reviews about products on the internet											
1	1.2	5	6.2	14	17.3	35	43.2	26	32.1	3.99	0.93
30. I am happy to use my credit card to purchase from an online vendor											
3	3.8	14	17.5	18	22.5	29	36.3	16	20.0	3.51	1.11
Customer's trust in used online shopping websites										3.83	0.54

Table Ⅲ and figure (9, 10) show that the total of means of the variable customer's trust in used online shopping websites was 3.83 (SD0.54), which means that trust was low. Also,Q7"The websites on the internet enable me to search and buy materials faster" had the highest mean (4.23, SD0.83), while Q30"I am happy to use my credit card to purchase from an online vendor" had the lowest (3.51, SD1.11).

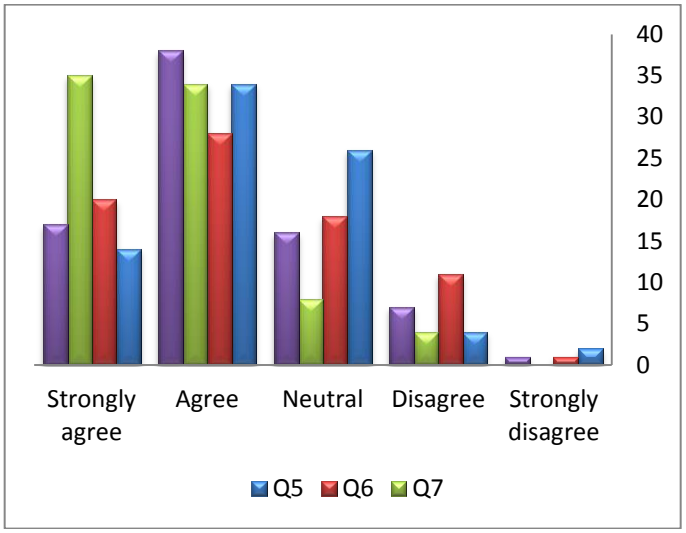


Fig. 9. Customers' trust in used online shopping websites

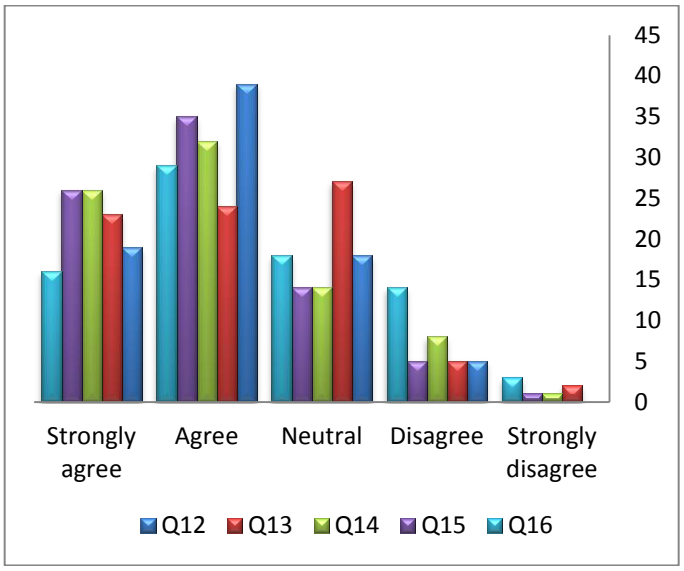


Fig. 10. Customers' trust in used online shopping websites

TABLE IV. SECTION 3 – EFFECT OF FORUMS, COMMUNITIES AND SOCIAL NETWORKING WEBSITES ON CUSTOMERS' ONLINE SHOPPING EXPERIENCE

Strongly agree		**Agree**		**Neutral**		**Disagree**		**Strongly disagree**		**Mean**	**SD**
N	%	N	%	N	%	N	%	N	%		
9. I am familiar with inquiring about material ratings on the internet											
2	2.6	5	6.5	22	28.6	25	32.5	23	29.9	3.81	1.03
10. I trust my friends on online forums and communities											
3	3.8	12	15.0	13	16.3	31	38.8	21	26.3	3.69	1.13
11. I use online forums and communities for acquiring information about a product											
2	2.6	6	7.7	16	20.5	32	41.0	22	28.2	3.85	1.01
Effect of forums, communities and social networking websites on customers' online shopping experience										3.78	.790

Table VI and figure 11 show that the total of means of the variable effect of forums, communities and social networking websites on customers' online shopping experience was 3.78 (SD0.79), which means that this effect was low. Also,Q11"I use online forums and communities for acquiring information about a product" had the highest mean (3.85, SD1.01), while Q10"I trust my friends on online forums and communities" had the lowest (3.69, SD1.13).

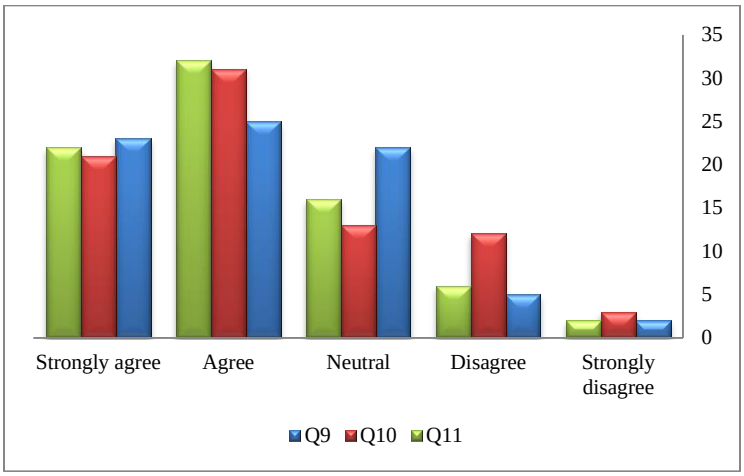


Fig. 11. Effect of forums, communities and social networking websites on customers' online shopping experience

TABLE V. SECTION FOUR – LEARNING TO USE COMPUTERS AND INTERNET FOR SHOPPING

Strongly agree		**Agree**		**Neutral**		**Disagree**		**Strongly disagree**		**Mean**	**SD**
N	%	N	%	N	%	N	%	N	%		
16. It is easy to become skilful at using the websites											
3	3.8	2	2.5	10	12.5	36	45.0	29	36.3	4.07	0.97
17. Learning to operate the websites on the internet is easy											
1	1.3	4	5.1	1	1.3	30	38.0	43	54.4	4.39	0.85
27. I have had training to use computers and the internet											
3	3.8	13	16.7	12	15.4	24	30.8	26	33.3	3.73	1.20
28. I have learned to use the internet to shop online											
1	1.3	14	18.2	10	13.0	28	36.4	24	31.2	3.78	1.12
29. My learning and training is/was useful for online shopping											
1	1.3	7	9.1	10	13.0	30	39.0	29	37.7	4.03	1.00
Examine learning to use computers and internet for shopping										4.00	0.699

Table V and figure 12 show that the total of means of the variable examine learning to use computers and internet for shopping was 4.00 (SD0.699), which indicates a low level of learning to use computers and internet for shopping. Also, Q17"Learning to operate the websites on the internet is easy" had the highest mean (4.39, SD0.85), while Q27"I have had training to use computers and the internet" had the lowest (3.73, SD1.20).

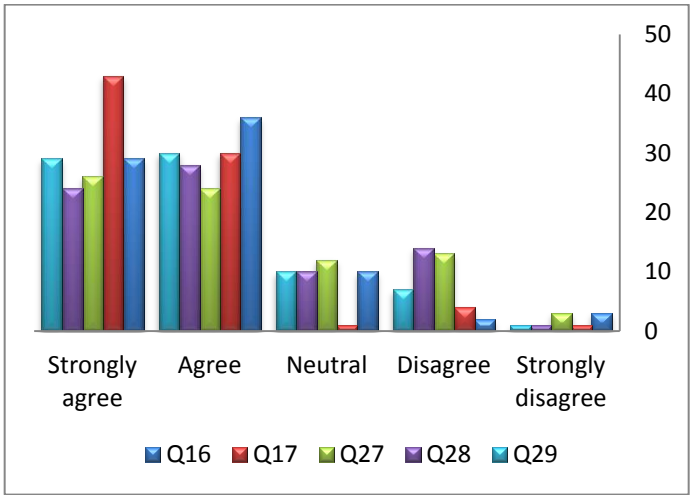


Fig. 12. Learning to use computers and internet for shopping

TABLE VI. SECTION FIVE – INTERACTING WITH VENDER VIA ONLINE SHOPPING WEBSITES

Strongly agree		Agree		Neutral		Disagree		Strongly disagree		Mean	SD
N	%	N	%	N	%	N	%	N	%		
19. The website that I use for my online shopping is flexible to interact with											
3	3.8	1	1.3	12	15.4	30	38.5	32	41.0	4.12	0.98
20. My interaction with the websites in the internet is clear and understandable											
1	1.3	5	6.3	17	21.3	38	47.5	19	23.8	3.86	0.90
21. There is a sense of human warmth in the website that I use for my online shopping											
5	6.4	6	7.7	16	20.5	27	34.6	24	30.8	3.76	1.16
22. I perceive myself pretty experienced in using the computer											
3	4.0	2	2.7	6	8.0	40	53.3	24	32.0	4.07	0.94
23. I perceive myself pretty experienced in using the internet											
2	2.6	3	3.9	12	15.8	26	34.2	33	43.4	4.12	0.99
24. I have been using the internet for a long time											
3	3.8	7	8.9	10	12.7	31	39.2	28	35.4	3.94	1.09
25. There is a sense of human contact in the website											
0	.0	6	7.5	18	22.5	35	43.8	21	26.3	3.89	0.89
26. There is a sense of sociability in the website that I use for internet shopping											
1	1.3	9	11.4	16	20.3	37	46.8	16	20.3	3.73	0.96
Examine interacting with vender via online shopping websites										3.92	0.64

Table Ⅵ and figure (13, 14) show that the total of means of the variable examine interacting with vender via online shopping websites was 3.92 (SD0.64), which indicates a low level of such behavior. Also, Q19 "The website that I use for my online shopping is flexible to interact with" and (23) "I perceive myself pretty experienced in using the internet" had the highest mean (4.12, SD0.98, 0.99 respectively), while Q26"There is a sense of sociability in the website that I use for internet shopping" had the lowest (3.73, SD0.96).

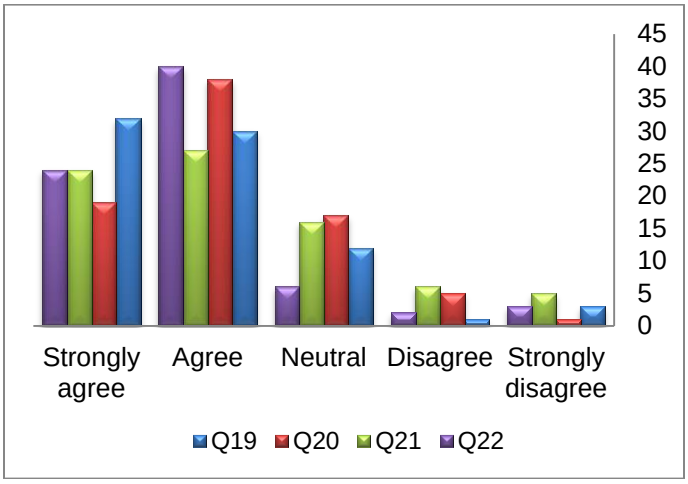


Fig. 13. Interacting with vender via online shopping websites, Q19-22

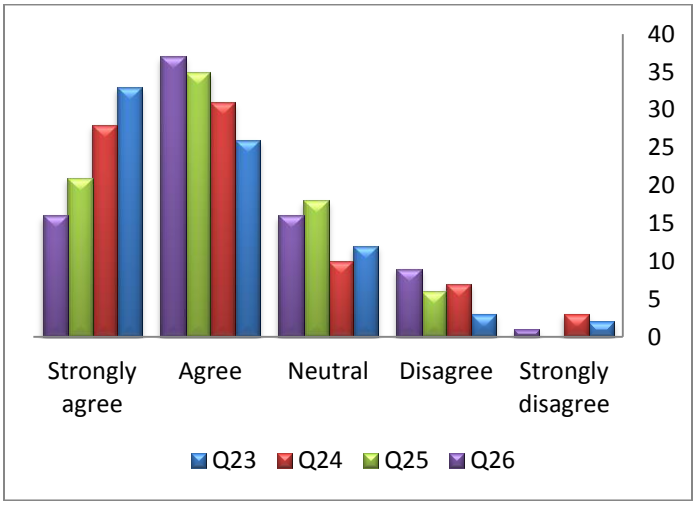


Fig. 14. Interacting with vender via online shopping website, Q23-26

TABLE VII. MEANS, SDS AND T-TEST FOR THE GENDER OF THE MEMBERS OF THE STUDY SAMPLE ON THE SECTIONS

Gender	N	Mean	SD	Std. error mean	t	df	Sig. (2-tailed)
Ability to search about desired product online							
M	38	4.06	.627	.102	1.375	79	.173
F	43	3.87	.623	.095	1.374	77.638	.173
Customer's trust in used online shopping websites							
M	38	3.86	.572	.093	.494	79	.623
F	43	3.80	.512	.078	.491	74.866	.625
Effect of forums, communities and social networking websites on customers' online shopping experience							
M	38	3.82	.777	.126	.426	79	.672
F	43	3.74	.813	.124	.427	78.510	.671
Examine learning to use computers and internet for shopping							
M	38	4.07	.807	.131	.822	79	.413
F	43	3.94	.590	.090	.807	67.002	.423
Examine interacting with vender via online shopping websites							
M	37	3.90	.727	.120	-.321-	78	.749
F	43	3.94	.554	.084	-.315-	66.659	.754

Table Ⅶ and figure 15 show that the significance value for all sections is greater than, 05, which means that there are no

statistically significant differences in all sections between males and females.

V. Critical Evaluation of Results

As stated in research methodology section, a questionnaire was distributed in several branches of PGS. The questionnaire was filled by random customers who attend this shop randomly and voluntary. These customers represent a population of Jordanian society and their opinion towards E-commerce adoption that is generally representative for other SMEs in the HKJ. Nevertheless, the insights of this mainly concerned the particular vision of PGS customers' enthusiasm for their vendor's migration into E-commerce.

The results that obtained from the survey conducted showed that 57.4%of participants had no online shopping experience, while 42.6%had some online experience. Based on literature review studies about E-commerce in Jordan, previous studies identified the impacts of society culture, with results generally affirming more online shoppers than expected, despite the entrenched Jordanian societal cultural preference for traditional shopping formats.

However, the fact that the survey was conducted in Amman, the capital of HKJ, is an important consideration. Amman represents the most prosperous city in the country where all required services are available, including education and other facilities (e.g. internet infrastructure). Furthermore, most of the respondents are relatively recent graduates(74.7%) aged 20-25 years old (37.8%),representing the most computer literate segment of Jordanian society; older and less educated portions of the population – most of whom would benefit most from E-commerce applications –are not represented. The majority of participants (66.7%) were females, but no gender effect was determined by the t-test of the study samples in all sections.

The most striking finding of the study is that the majority of respondents did not have direct online shopping experience, which demonstrates the fact that Jordanian societal culture has a major effect on organizations' intention to move towards E-commerce; put simply, organizations do not seriously contemplate developing their E-commerce activities, and the majority of people in local society are unresponsive to such initiatives.

Generally, HKJ as previously stated in literature review, is a developing country with limited resources that needs enhancement to accept the E-commerce adoption. Given the lack of interest among the public and SMEs, the government must play a more active role in facilitating E-commerce with appropriate infrastructure and expanding E-commerce knowledge to people across the country. This may be spearheaded by the full implementation of the ambition E-government project in the country, which may encourage more people across the country to gain more knowledge of E-commerce potential based on their experience of the convenience of E-government facilities for taxation, public services fees and transaction payments online, circumventing the cumbersome bureaucracy of traditional interactions with the government. This will erode Jordanians' fears of local organizations' E-commerce adoption and increase their enthusiasm about E-commerce generally, promoting the development of serious strategies to apply and benefit from this change.

Local E-commerce companies should be actively encouraged and motivated by incentives, such as the elimination of sales tax for E-commerce transactions for a probationary period in order to encourage private sector companies to conduct serious steps towards adopting E-commerce.

Finally, based upon previous researches conducted on E-commerce adoption in Jordan and results of this research, it highly recommended that PGS in particular wait a period of time (i.e. a minimum of two years) to begin changes preparatory for the serious adoption of E-commerce in their business operations due to market uncertainties at the present juncture.

VI. Conclusions

It can be concluded that Jordanian society is strongly attached to traditional ways of shopping, but the upcoming generations, represented by fresh graduates in this study, are increasingly experienced and accepting of the potential of online shopping experiences. This means that changes in consumer behavior (i.e. shopping trends) can be anticipated in the near future, with the diffusion of more advanced computer literacy throughout Jordanian society. Furthermore, increased usages of SNS in Jordanian society in daily life is affecting shopping behavior regardless of age category.

In conclusion, SMEs are encouraged to prepare a plan, design and implement the migration towards adopting full E-commerce activities to encompass the development of Jordanian customers' acceptance of online shopping.

References

[1] P. Permwanichagun, S. Kaenmanee, and A. Naipinit, "The External Environments Factors Affecting Success For Implementation: In Context Of Sole Proprietorship E-Commerce Entrepreneurs In Thailand". International Business Management, Vol. 9, pp. 122-127. 2014

[2] M. Aljaber, "The impact of privacy regulations on the development of electronic commerce in Jordan and the UK", Doctorat thesis, De Montfort University, UK, 2012.

[3] S. Mariotti, F. Sgobbi,"Alternative paths for the growth of e-commerce". Futures 33(2):109-125 · March 2001

[4] M. Yasin, M., Alavi, J., Czuchry, A. and Shafieyoun, R.,"An exploratory investigation of factors shaping electronic commerce practices in Iran: Benchmarking the role of technology and culture", *Benchmarking: An International Journal*, Vol. *21*(5), pp.775-791, 2014.

[5] S. Allahawiah, H. Altarawneh, and S. Alamro "The Internet and Small Medium-Sized Enterprises (SMES) in Jordan". World academy of Science, Engineering and technology,Vol. 62, pp.302-306, 2010.

[6] N. Terzi,"The impact of e-commerce on international trade and employment". *Procedia-Social and Behavioral Sciences*, Vol. 24, pp.745-753. 2011.

[7] M. Khoshnampour and M. Nosrati, "An overview of E-commerce".*World Applied Programming*, Vol. *1*(2), pp.94-99. 2011.

[8] R. Zaker and A. Ansari, "Towards Improving Quality of E-Commerce Websites in Hospitals". *Istanbul, Turkey MAY 8-10, 2013.*

[9] Z. Huang, and M. Benyoucef, " From e-commerce to social commerce: A close look at design features". *Electronic Commerce Research and Applications, Vol. 12*(4), pp.246-259. 2013.

[10] E. Turban, D. King, J. Lee, T. Liang, and D. Turban, "*Electronic commerce: A managerial and social networks perspective"*. Springer.2015.

[11] G. Nielson, B. Pasternack, A. Viscio, "Up the (E) organization! A seven-dimensional model for the centerless enterprise". Managing Mag [e-journal], *Vol*. 18, pp.52– 61 (First Quarter). 2000.

[12] W. Hong, K. Zhu, "Migrating to internet-based e-commerce: Factors affecting E-commerce adoption and migration at the firm level". Information & Management [e-journal], *Vol*. 43, pp.204–221. 2006.

[13] S. Lippert, and C. Govindarajulu, "Technological, organizational, and environmental antecedents to web services adoption". *Communications of the IIMA*, *Vol*. *6*(1), p.14. 2015.

[14] R. Lituchy, "Bed and breakfasts, small inns, and the internet: the impact of technology on the globalization of small businesses". Journal of International Marketing, *Vol*. 8(2), pp. 86–97. 2000.

[15] J. George, "Influences on the intent to make Internet purchases". *Internet Research-Electronic Networking Applications and Polic*y, *Vol*. 12, pp. 165-180. 2002.

[16] A. Charlesworth, "Digital marketing: A practical approach". Routledge, 2014.

[17] S. Drew, "E-Business research practice: towards an agenda". *Electronic Journal on Business Research Methods Vol.* 1(1). *pp.18-26.* 2002.

[18] R. Stair, and G. Reynolds, "Principles of Information Systems". (Third Edition). Course Technology a division of Thomson Learning, Inc: Canada. 2003.

[19] M. Abbad, R. Abbad, and M. Saleh, "Limitations of e-commerce in developing countries: Jordan case", Education, Business and Society: Contemporary Middle Eastern Issues, Vol. 4 Iss: 4, pp.280 – 291, 2011.

[20] M. Al-debei, "The current state of e-commerce in Jordan: applicability and future prospects "an empirical study"". *University of Jordan e-library.2005.*

[21] S. Alamro, S. Tarawneh, "Factors Affecting E-Commerce Adoption in Jordanian SMEs. European Journal of Scientific Research. Vol. 64(4), pp.497-506.. 2011.

[22] A. Al-Bakri, M. Katsioloudes,"The factors affecting e-commerce adoption by Jordanian SMEs", Management Research Review, Vol. 38 Iss: 7, pp.726 – 749, 2015.

[23] M. Hajli, H. Bugshan, M. Hajli, and A. Kalantari, "E-Commerce Pre-Adoption Model for SMEs in Developing Countries". In Proceedings of the 2012 International Conference on e-Learning, e-Business, Enterprise Information Systems, and e-Government, Las Vegas, United States.2012.

Wavelet based Scalable Edge Detector

Imran Touqir and Adil Masood Siddique
Military College of Signals
National University of Sciences and Technology (NUST)
Islamabad, Pakistan

Yasir Saleem
Electrical engineering Department
University of Engineering and Technology (UET)
Lahore, Pakistan

***Abstract*—Fixed size kernels are used to extract differential structure of images. Increasing the kernal size reduces the localization accuracy and noise along with increase in computational complexity. The computational cost of edge extraction is related to the image resolution or scale. In this paper wavelet scale correlation for edge detection along with scalability in edge detector has been envisaged. The image is decomposed according to its resolution, structural parameters and noise level by multilevel wavelet decomposition using Quadrature Mirror Filters (QMF). The property that image structural information is preserved at each decomposition level whereas noise is partially reduced within subbands, is being exploited. An innovative wavelet synthesis approach is conceived based on scale correlation of the concordant detail bands such that the reconstructed image fabricates an edge map of the image. Although this technique falls short to spot few edge pixels at contours but the results are better than the classical operators in noisy scenario and noise elimination is significant in the edge maps keeping default threshold constraint.**

Keywords*—Wavelet scales correlation; Edge detection; image denoising; Multiresolution analysis; entropy reduction

I. Introduction

Spatial domain, frequency domain and wavelet based techniques are being used independently to detect edges in an image. Spatial filters are good at localization accuracy but lack control over the operator's scale. Similarly Fourier transform being global in nature can neither localize sharp transients nor differentiate between true and false edges under noisy scenario. The classical edge detectors [1,2] do not yield adequate edge maps of the noisy images over default threshold values. The choice of optimum threshold for edge detection [3] is not generic. A good threshold assigned to yield a good edge map for a particular type of image and noise model may be inappropriate for other type of image or the different noise model. Thus it requires user's intervention to assign suitable threshold value to differentiate between true and false edges. A multiscale edge detection algorithm has been presented in [4] for SAR images but it is not advocated for the low PSNR images Thus the two major dilemmas for edge detection are; firstly the choice of appropriate threshold [5] to segregate noise and true edges and secondly to opt for an appropriate scale for edge detection.

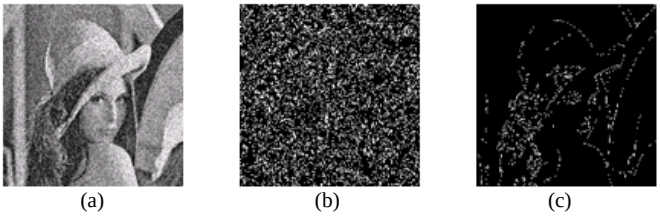
Usually threshold is empirically found using trial and error process and varies for different noise models and intensities in the image. Figure 1 highlights the results of edge detection using default and manually assigned threshold value for a noisy image.

Another dilemma in edge detection is that the edge operators are fixed size masks. Compactly supported kernels are good to identify sharp transients whereas fall short to spot structural variations in the image. On the other hand large size masks are good to identify large scale variations but these are not sensitive to swift variations and lose localization accuracy and fidelity. Thus the objective of this paper is to explore an edge detection paradigm such that:

- It efficiently works on default threshold value, does not require user intervention to assign an appropriate threshold value and thus can be independently used in any pre-processing stage in digital image processing applications.
- It incorporates an inbuilt technique for partial noise elimination that holds equally for different noise models and intensities.
- It facilitates scalability in edge detection.

II. Noise Models in Images

The noise models assume that the noise is oscillatory and image is smooth or piecewise smooth. Segregation of noise and information in a signal is an ill posed problem. High noise fluctuates image entropy hiding information contents of the image and behaves differently for versatile images [6],[7]. More information an image has more abruptly entropy maxima will be reached by noise induction. Figure 2(a) depicts that image entropy is proportional to noise induction and it decreases monotonically with increase of noise after its maxima.



(a) (b) (c)

Fig. 1. Effects of threshold on edge detection. (a) Noisy Lena image. (b) Edges detected by Canny using default threshold. (c) Edges detected by Canny using threshold as 0.37

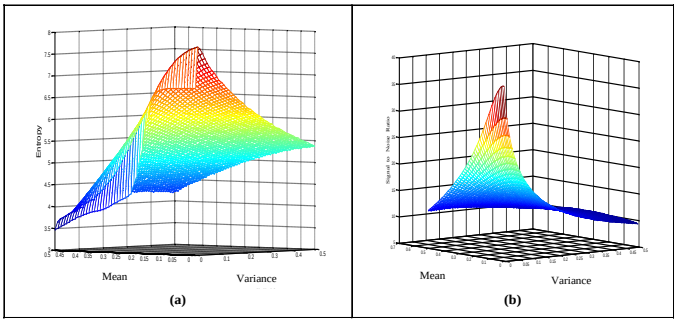


Fig. 2. Image entropy (a) PSNR for Gaussian noise induced in Lena image for varying mean and variance. Figure 2(b) depicts image PSNR under the influence of Gaussian noise for varying noise parameters; used to evaluate the proposed algorithmic validity

III. Prepare Scalability in Edge Detection

Multiresolution analysis (MRA) is concerned with the representation and analysis of signals or images at more than one resolution [8-10]. The appeal of such an approach is obvious; the features that might go undetected at one resolution may be easy to spot at another. Wavelet analysis is well suited to isolate sharp transients in a signal, a task at which Fourier analysis is not so pleasing. Analysis of images with QMF has been exploited for edge detection. For images, a two dimensional scaling function $\varphi(x, y)$ and three two dimensional wavelets, $\psi^1(x, y)$, $\psi^2(x, y)$ and $\psi^3(x,y)$ are required.

$$\varphi(x,y) = \varphi(x)\varphi(y) \quad (1)$$

$$\psi^1(x,y) = \psi(x)\varphi(y) \quad (2)$$

$$\psi^2(x,y) = \varphi(x)\psi(y) \quad (3)$$

$$\psi^3(x,y) = \psi(x)\psi(y) \quad (4)$$

Equation (1) calculates the approximation and remaining (2) to (4) calculate gradients along horizontal, vertical and diagonal directions respectively.

Image resolution ascertains the choice of appropriate scale [11] for edge detection which is not adjustable with classical edge detectors. However with the wavelet model, we can construct our own edge detector with appropriate scale. Scale is controlled by regularization parameters that further control the significance of edges to be shown. Edges of higher significance are likely to be kept by the wavelet transform across scales and lower significance are likely to disappear when scale increases [11]. Wavelet filters of large scales are more effective for removing noise, but at the same time increase the uncertainty of the edges locations. Small scale wavelet filters have good localization accuracy, but can hardly distinguish between noise and true edges. Many techniques have been proposed for multiscale edge detection [12,13], however, there is less agreement on the following;

- Number of scales of edge detector or decomposition levels.
- Methods to opt for optimum scale.
- How to synthesize the results at different scales.
- Choice of threshold value.

Edge detection based on wavelet analysis is efficient in the sense that it requires least visual interpretation. Different wavelet basis functions have different waveforms, central frequencies and vanishing moments. Suitable decomposition level is desirable to maintain a clear background, edge contour and to remove irrelevant higher frequency components on the surface. The decomposition performed by different wavelet function captures features with different spatial frequencies based on the characteristics of the selected wavelet function at each level. Theoretically, the wavelet decomposition can be iterated n times on an image, where $2n \leq m < 2n+1$ and m is the min of number of pixels of an image in either direction. Image decomposition up to apex is a non-optimal solution for edge detection. Total bands constituting directional edges are thrice the decomposition level with an additional approximation band which is susceptible to further wavelet decomposition. Decomposition of lower resolution generates artifacts and discontinuity in edges. Similarly thick edges support high level decomposition where as thin edges in images suffer more edge losses as scale increase. Further that noise in the image is inversely proportional to decomposition level for edge detection. Due to vast diversity and complexity in the image structural parameters and noise models/ intensities, optimal decomposition level (n) for edge detection has not been derived. However its dependence on three image parameters i.e. resolution(r), structural parameters(s) that includes statistical parameters and noise level(η) has been established.

$$n = f(r, s, \eta) \quad (5)$$

IV. Wavelet Synthesis for Edge Detection

The lower resolution wavelet detail bands are interpolated to the original image size that partially recaptures the missing edge pixels besides facilitating matrix multiplications of the concordant wavelet bands. Equations (1)-(4) can be exploited by WSC to detect edges from an image in nine steps S-1 to S-9:

S-1. A pair of QMF is operated on gray level image in vertical followed by horizontal direction.

S-2. Decimation by two after each filtering stage is applied and high frequency details are extracted at level-1.

S-3. On the magnitude image so obtained thresholding is performed to obtain the edge map at level-1. Default threshold is taken as one fourth of the band mean value of the wavelet coefficients.

S-4. The coefficient values outside three sigma range in the approximation bands are chopped off to three sigma values.

S-5. The resultant lowpass residue is taken for analysis to get second level decomposition. Steps S-1 to S-4 are repeated to obtain edges at level 2.

S-6. Lowpass residue is carried over from previous level to iterate up to nth level. Edge details of different precision are obtained at each decomposition level.

S-7. The inbuilt noise suppression technique and down sampling diminishes few edge pixels. The lower resolution bands are interpolated by nearest

neighborhood up to original image size that facilitates capturing of few faded edge pixels and matrix multiplication.

S-8. The horizontal, vertical and diagonal interpolated bands up to nth level are point wise multiplied respectively and the product of concordant detail bands are cumulated. The synthesis of product bands if yielded at level-1 is re-interpolated to match the original size of the image. However it is convenient to interpolate the product bands to the original size and then synthesize.

S-9. The harmonic mean of the cumulated detail band yields image edge map.

Figure 3 shows the practical implementation of proposed algorithm showing the wavelet analysis filter bank for the input image I followed by its synthesis filter bank for edge detection for n=4. E is the resultant image edge map. The optimum decomposition level is not generic. Initially [6,7] results were compiled up to fourth level wavelet decomposition which are enhanced to nth level based on subjective analysis. The analytical expression for WSC edge detection is as follows:

$$\psi^1 = \prod_{i=1}^{n} \psi_l^1 \tag{6}$$

$$\psi^2 = \prod_{i=1}^{n} \psi_l^2 \tag{7}$$

$$\psi^3 = \prod_{i=1}^{n} \psi_l^3 \tag{8}$$

$$\psi = \sum_{d=1}^{3} \psi^d \tag{9}$$

$$E = \sqrt[n]{\psi} = \left(\sum_{d=1}^{3} \prod_{i=1}^{n} \psi_i^d \right)^{1/n} \tag{10}$$

Where ψ represents the synthesis of all the detail band coefficients by the given technique, superscript d=1,2,3 represents interpolated horizontal, vertical and diagonal detail bands to the original image size after multiplication of concordant bands, subscripts l represents the decomposition level. E is the resultant edge map of the image, n is the decomposition level and its dependence has been established in (5).

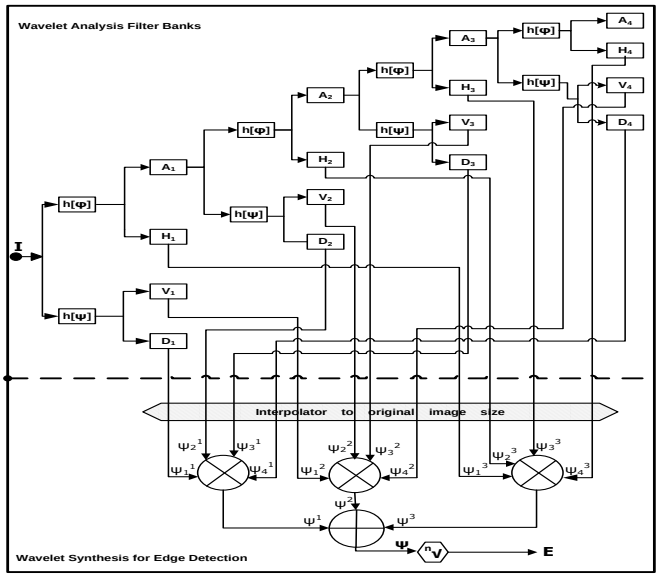


Fig. 3. Wavelet analysis filter banks and its synthesis for edge detection for n=4

A, H, V and D represents decimated approximations, horizontal, vertical and diagonal detail coefficients, subscripts indicate decomposition level. ψ1, ψ2 and ψ3 represents horizontal, vertical and diagonal interpolated edge maps after point wise multiplication of concordant band coefficients up to nth level respectively. Encircled X and + denote multiplier and summer respectively. E is the resultant image edge map.

The computational complexity of the algorithm is linear. It is proportional to the order of the wavelet filter, decomposition level and the interpolation technique used for synthesis. Haar offers least complexity. The computational complexity for first level wavelet decomposition using Haar equals to N and for nth level becomes $N^2 \sum_{l=0}^{n} 2^{2(1-l)}$. During WSC synthesis the image in lower resolutions are interpolated to original size prior to multiplication of the concordant bands. The nearest neighborhood interpolation in 2D is the order of N2 comparisons. Bilinear interpolation consists of N2 multiplications. The details bands up to nth level are interpolated. The complexity of each directional detail band becomes n x N2 and total interpolation complexity turns out to be 3 x n x N2. The harmonic mean of the image prior to display adds complexity by N2. Thus the overall complexity of the algorithm for (10) is

$$N^2\left\{\left(\sum_{l=0}^{n} 2^{2(1-l)}\right)+3l+1\right\} \quad (11)$$

The computational complexity of image size 512 x 512 for edge detection at 4th level is equal to 3.6599x106.If the n^{th} root of product of interpolated approximation bands up to nth level is added to the edge detected image, it gives denoised image with reduced entropy.

$$R=\left(\prod_{i=1}^{n}\varphi_l+\sum_{d=1}^{3}\prod_{i=1}^{n}\psi_i^d\right)^{1/n} \quad (12)$$

where R is the denoised reconstructed image with reduced entropy. The product of approximation concordant band's computational complexity is added in (11) by a factor of 4nN2. However analysis of denoised reconstructed image is not carried out in this paper.

V. Quality Metric

The pixels constituting edges are delocalized; therefore legitimate MSE does not correlate with Psycho visual comparison. The absolute difference in the Distance Transform (DT) [14] of edges detected from original and noisy image is taken as measure of error. PSNR based on the DT [5],[6] is evaluated as

$$PSNR=10Log_{10}\frac{\chi}{\|DT_1-DT_2\|_2^2} \quad (13)$$

where χ is the peak signal value which is 255 in the experiments. DT1 is the DT of the edge detected image from original image and DT2 is the DT of the edge detected image from noisy image and their second norm is taken that computes MSE. If m and n are the rows and columns of the DT matrix respectively such that dt1(m,n)∈DT1 and dt2(m,n)∈DT2 then their second norm is defined as

$$\|DT_1-DT_2\|_2^2=\frac{1}{m\times n}\sum_{m}\sum_{n}|dt_1(m,n)-dt_2(m,n))| \quad (14)$$

Entropy H(x) of the edge map is determined as

$$H(X)=-\sum_{i}P(x_i)\ln(x_i) \quad (15)$$

where P(xi) is the probability of ith pixel value. It is not possible exactly to infer the entropy measure for edge map of the image because the entropy variation in the image due to noise or information contents is an ill posed problem. There exists no such known method to infer weather the increase or decrease in entropy of the image edge map is due to variations in noise density or there exists true edges. Among different wavelet basis functions, the entropy values obtained at optimal decomposition levels will be different, which means that their information contents will also be different. In this sense, the higher entropy value is where more information is contained. The entropy criteria used coincide with [15],[16] within the family of wavelets and supplements psycho visual comparison.

VI. Experimental Results

The edges detected by WSC and Canny edge detector for Lena image of resolution 512 x 512 and 128 x 128 as demonstrated in Figure 4 has yielded better results for high resolution images. Classical detectors failed to extract edges from Lena image with N(0, .02) (Figure 4d) and were dominated by noisy pixels (Figure 4e) on default threshold values where as proposed scheme yielded significant edge map of the image (Figure 4f) keeping the same noise level. Thick edges are vulnerable at multiple scales, thus are prominent in the final edge map of the image. Further that Canny failed to produce edge map of Lena at default threshold for N(0, .02) or above. Whereas proposed scheme has given an adequate edge map of Lena upto noise of variance 0.09 as shown in Figure 4(g, h, i). Figure 4(j, k, l) also reveals that the proposed scheme does not give equivalent results for low resolution images and for Lena image of resolution 128 x 128 Canny performed better than the proposed technique. The proposed detector is equally good for other noise models as well. Results for uniform noise induced in the image are trivial due to wavelets in built approximating and detailing characteristics.

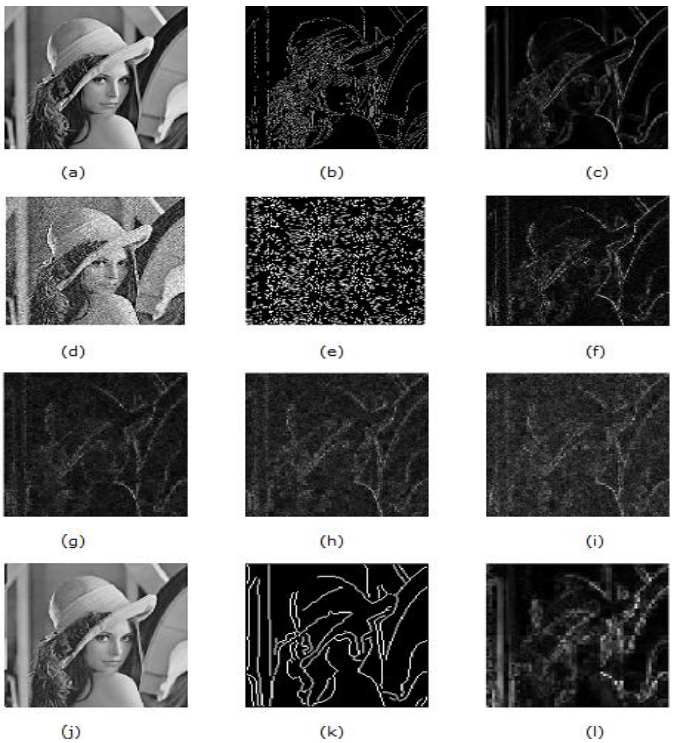


Fig. 4. Edge detection comparison using Lena image over default threshold value. (a) Lena image of resolution 512x512 (b) Canny edge detector (c) edges detected by WSC (d) Gaussian noise N(0,.02) induced in Lena image (512x512), (e) Canny edge detector result (f) edges detected by WSC from (d). Similarly (g), (h) and (i), are edge maps by the proposed technique of Lena image(512x512) induced with Gaussian noise of variance .04, .05 and .09 respectively. (J) Lena image of resolution 128 x 128, (k) is edge detected by Canny and (l) is the proposed edge detector

Figure 5 shows Lena image subjected to salt and pepper noise with noise density .01, .02 and .05. The edge detection by classical filters are dominated by spikes and unable to classify

true edges whereas isolated spikes have been suppressed by the said technique. Significant edge map of the image has been obtained for noise density up to 0.05. Canny clutters the results with false edge points where as due to scale multiplication of concordant band coefficients, the edges which are revealed in the image structure and present at multiple scales are captured in proposed scheme. Remarkable difference can be seen between classical edge detector 5(e) and the proposed edge detector 5(f). It even yielded significant edge map for noise variance as high as 0.09. Similarly Figure 6 supports the preceding results using Boat image.

The results of the proposed technique also depend upon the image structure which includes edge thickness and the edge quality varies for different images at same noise level. The strength of the algorithm is such that it works for diverse images on default threshold values without user's intervention for operating parameters.Db1 gave optimum results within the family of wavelets. The increase in the length of wavelet filter in Daubechies family increases the number of vanishing moments that blurs the edges. Further comparison of natural and synthetic images for edge detection for different noise models exports similar results. Although the results are inferior to DSCED and DSCANED edge detectors [17] but these do not have standard parameters and require user's intervention during the edge detection for assigning suitable parameters for optimum results.

Db1 scale correlation furnished most favorable detection within the conducted experiments due to its compact support. The edge blurring occurs with increase of the length of wavelet filter coefficients. Entropy of spatial domain filters decrease strictly monotonically with increase of noise variance. Experimental results reveal that it is intricate to distinguish information and noise contents in an image by the classical edge detectors. However exploiting correlation at different resolutions, structural details are retained coupled with noise suppression. The image entropy variations under the influence of Gaussian noise are function of amount of information in the image, its intensity values and the noise model. DWT filters preserve more information and fluctuates around 6 bits. Difference of entropies of spatial domain [1-3] filters and wavelet filters is eminent in Figure 7. DWT level-1 edge detector has the maximum entropy followed by level -2, level-3, level-4 and the proposed algorithm. Entropy of the proposed algorithm is decreased due to partial noise suppression. The noise saturates pixel's intensity values and conceals intelligence contents of the image. The entropy maxima by the increase in noise density changes from image to image. Greater the information contents an image has more suddenly the maxima will be reached and vice versa. In spatial domain filters maximum entropy is preserved by Canny. The optimal results of Canny depend upon selection of optimal threshold for edge detection, however, in this work all the experimental results are based on default threshold values.

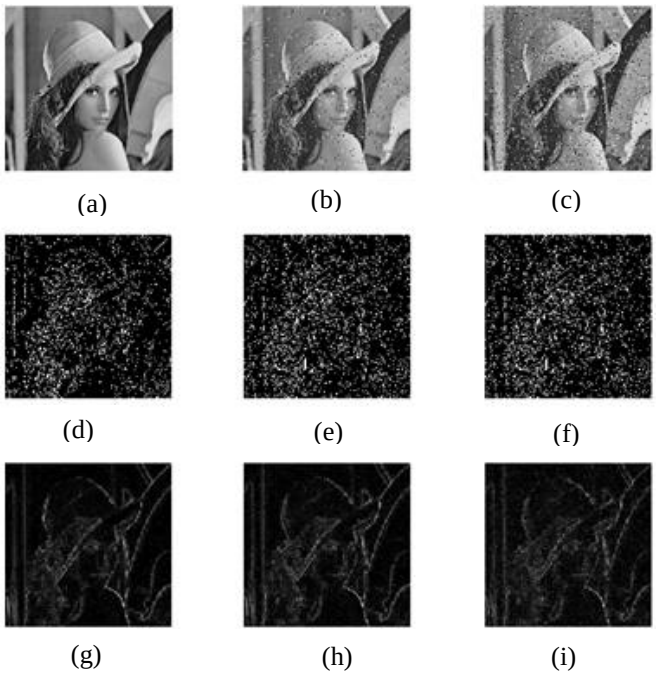


Fig. 5. Salt & pepper noise induced in Lena image with density (a) 0.01, (b) 0.02, (c) 0.05, (d), (e), and (f) are edges detected by Canny respectively and (g), (h), (i) are the edges detected by proposed detector from (a), (b), and (c) respectively

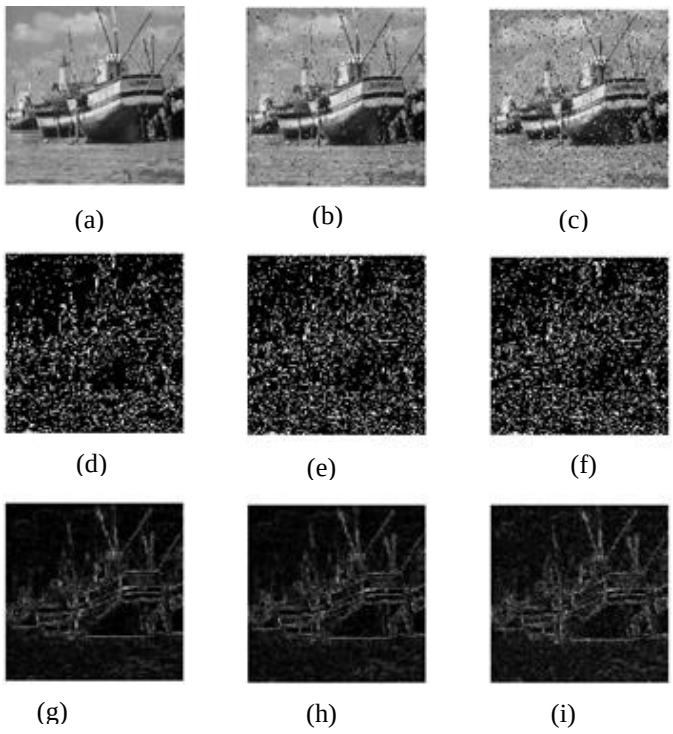


Fig. 6. Salt & pepper noise induced in Boat image with density (a) 0.01, (b) 0.02, (c) 0.05, (d), (e), and (f) are edges detected by Canny respectively and (g), (h), (i) are the edges detected by proposed detector from (a), (b), and (c) respectively

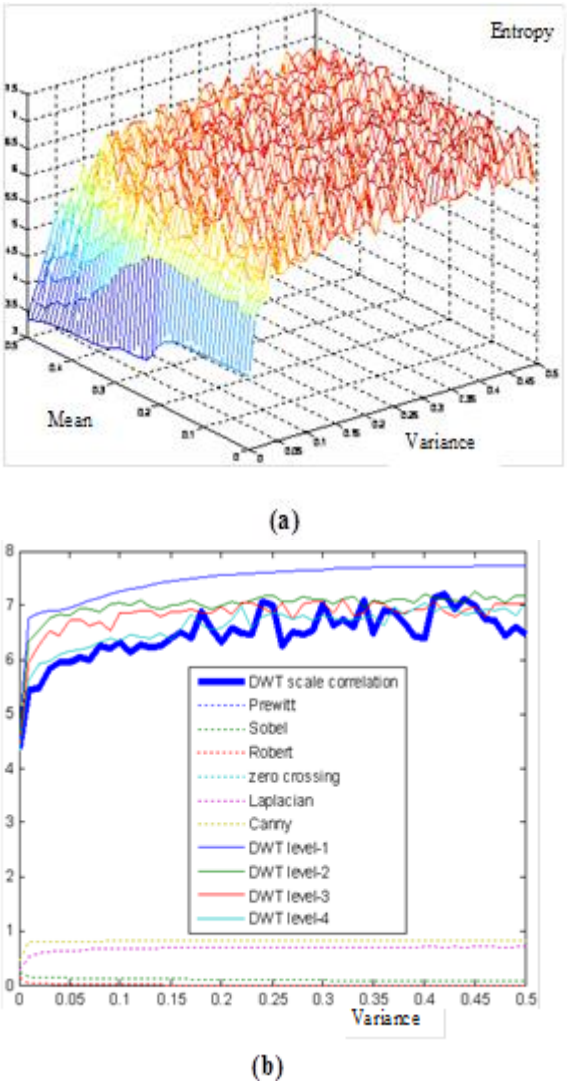


Fig. 7. (a) Entropy of edges detected from Lena image under Gaussian noise of varying mean and variance. (b) Comparison of entropies for different edge detectors for Gaussian noise of varying variance

VII. CONCLUSION

The scale correlation of concordant wavelet bands significantly capture edges present at multiple scales and elegantly discard isolated spikes and thus noise is partially segregated from true edges. The technique justifies for higher resolution images. However wavelet decomposition level cannot be unique and varies from image to image and depends upon image resolution, type, noise model and pixel intensities. The image decomposition level is function of its resolution. The said technique has outperformed the existing schemes keeping default threshold constraint. The algorithm is equally applicable to images with depleted PSNR where conventional filters fall short to give adequate edge map. The algorithm is advocated for edge detection where noise model or noise intensity either varies or unpredictable prior to image processing. Greater the image noise level greater is the performance difference of the proposed detector with classical edge operators. Moreover it performs without the user's interaction and can be elegantly cascaded in preprocessing stage such as segmentation or feature extraction/matching. The reconstructed image through scale correlation gracefully suppresses noise, reduces image entropy and favors further processing in diverse image processing applications.

ACKNOWLEDGMENT

The research has been facilitated in image processing lab at Military College of Signals, National University of Sciences and Technology, Islamabad.

REFERENCES

[1] Rafael C. Gonzalez and Richard E. Woods. Digital Image Processing. 3rd ed. India: Pearson Education, 2008.

[2] Marr, D., Hildreth, E. Theory of edge detection. In: Proc. 1980 R. Soc. London. B, 29 February 1980, London. B, pp 187-217.

[3] Canny J. A computational approach to edge detection. IEEE Trans. PAMI 1986;8: 250-468.

[4] Tello Alonso M, Lopez-Martinez C, Mallorqui, J.J, Salembier P. Edge enhancement algorithm based on the wavelet transform for automatic edge detection in SAR images. IEEE Transactions on Geoscience and Remote Sensing 2011;49(1): 222-235.

[5] Wu, Y, He Y, and Cai H. Optimal threshold selection algorithm in edge detection based on wavelet transform. Image and Vision Computing 2005; 23(13):1159-1169.

[6] Muhammad Saleem, Imran Touqir and Adil Masood Siddiqui. Novel edge detection. In: IEEE 2007 International Conference on Information Technology (ITNG'07): 2-4 April 2007; Washington, DC, USA: IEEE Computer Society. pp. 175-180.

[7] Imran Touqir and Muhammad Saleem. Novel edge detector. Lecture Notes in Computer Science 2008; 4958: 432-443.

[8] Stephane G Mallat,. A theory for multiresolution signal decomposition the wavelet representation. IEEE Trans. on Pattern and Machine Intelligence 1989;11(7):674-693.

[9] Brain M. Sadler and A. Swami. Analysis of multiscale products for step detection and estimation. IEEE Trans. Information Theory 1999; 45:1043-1051.

[10] Park D. J, Nam K. N and Park R. H. Multiresolution edge detection techniques. Pattern Recognition Letters 1995;28(2):211-229.

[11] Touqir I and Saleem M. Novel wavelet synthesis for edge detection. Mehran University Research Journal of Engineering and Technology 2008;28(1):41-52.

[12] Ziou D. and Tabbone S. A multiscale edge detector. Pattern Recognition 1993;26(9):1305-1314.

[13] Junxi Sun, Dongbing Gu, Yazhu Chen, Su Zhang. A multiscale edge detection algorithm based on wavelet domain vector hidden Markov tree model. Pattern Recognition 2004;37(7):1315-1324.

Variability Management in Business-IT Alignment: MDA based Approach

Hanae Sbai[1] and Mounia Fredj[2]
[1,2] AlQualsadi Research team, ENSIAS, Mohammed V University of Rabat,
BP 713, Rabat 10000, Morocco

***Abstract*—The expansion of PAIS (Process Aware Information Systems) has created the need for reuse in business processes. In fact, companies are left with directories containing several variants of the same business processes, which differ according to their application context. Consequently, the development of PAIS has become increasingly expensive. Therefore, research in business process management domain introduced the concept of configurable process, with the aim of managing the variability of business process. However, with the emergence of the services-based development paradigm, the alignment of services with business processes is highly required in PAIS. Thus, in this paper an MDA based method which allows for generating configurable services from configurable process is proposed.**

Keywords—alignment; variability; MDA; PAIS; configurable service, configurable process

I. INTRODUCTION

Due to the lack of process control and automation into information systems centered data, the process orientation was established by the introduction of a new generation of information system called Process Aware Information System (PAIS), where the main unit of these information systems is the business process models. Thus, workflow management systems (WFMS) and integrated systems known as Enterprise Resource Planning (ERP) represent an example of a PAIS [1]. In the literature, a PAIS is considered as a software system that manages and executes business processes involving people, applications and information sources, based on a process model, while advocating separation of business logic and application logic [1].

In the last few years, with the wide adoption of PAIS, companies are left with directories containing several variants of the same business processes, which differ according to their application context. For instance, in the e-healthcare domain, 90 variants of "medical examination process" could be distinguished in a hospital [2]. Consequently, in order to choose or combine variants, the designer has to compare and adapt them manually, which could be a complex and an error prone operation. In this context, many research studies have focused on managing the variability of business processes by developing configurable processes [3] [4] [5] [6] [7] [8] [9].

Along with the improvement of business process reuse by the introduction of the variability management, the proposed approaches lack of business-IT alignment support. The study of existing works shows that these approaches do not allow the generation of configurable services emanating from configurable processes, in this context a new concept related to business processes has been introduced, which is the "service based process model" [14]. This has led us to study the alignment between the configurable processes and the enterprise applications, in particular services, with the aim of building PAIS that support service orientation. The emergence of variability management in business processes and services conduct the PAIS today to adopt the configurable processes at the business layer and the configurable services at the IT layer. In this perspective, an MDA (Model Driven Architecture) approach for the configurable services generation is developed in [15].

Therefore, it is argued that the alignment with supporting variability could also be beneficial. This alignment will enable the traceability management of business needs expressed at the business layer and their realization at the IT layer [16]. Indeed, it also allows change synchronization between the two layers. Consequently, the alignment is not limited only to establish the mapping between the configurable processes and configurable services, but also to maintain this correspondence when companies business needs evolve.

The paper is structured as follow: the concept of alignment supporting variability is firstly introduced and secondly the comparative study of different approaches is given. An MDA based method for configurable service generation is described in the Section IV.

II. BUSINESS-IT ALIGNMENT SUPPORTING VARIABILITY

A. Alignment concept

This section focuses on defining and discussing the most existing definitions of the alignment concept.

Alignment can be defined as the «dependency management» between business processes and services [18], «connection» of services to processes [19] or «change synchronization» between business processes and services [16].

According to the authors [20], the alignment of services with business processes is the ability to realize business process as a set of services. In this sense, the alignment allows for ensuring coherence between the processes of the business layer and services of the IT layer. It is considered that the alignment consists of the design of service-oriented architectures in a way that allows for easily adaptable business processes. This requires not only defining the dependency relationships between the activities of a business process and

related services, but also managing changes of processes and their related services.

In this work, business IT alignment concerns the generation of configurable services from configurable process in order to maintain consistency between the two layers (IT and business) and facilitate change management, taking into account the variability of processes and services.

Existing works on business IT alignment can be put under two main categories:

- Generation of services from a business process that involves designing business processes and defining the different rules that generate services from configurable process. These services will be implemented using the web services technology.
- Change management by analyzing the impact of changing the business process and services.

In this article, the main work is about the first category by focusing on the management of variability when it comes to generate services from a business process.

B. Service generation from business process

In this section, the concept of service generation is detailed in order to define the main elements which are required when it comes to generate services.

In the literature, the generation consists of transforming a business process into a set of services. This describes how services can be automatically generated from a business process.

The study of service generation works [14] [16] [17] [21] has shown that these works use the BPMN language [22] to represent a business process and the SoaML (Service Oriented Architecture Modeling language) [23] to represent the services to generate. For the generation, it is carried out in the most of time under the MDA (Model Driven Architecture) [24].

Furthermore, despite of the diversity of these works, there is no generation approach supporting the variability of the business process and services. In this perspective, an MDA based approach for the generation of VARSOAml configurable services from a Variant-Rich BPMN configurable process is proposed.

Thus, service generation with supporting variability requires:

a) Language for modeling a configurable process

b) Language for representing configurable services

c) MDA approach for configurable services generation

2) Modeling of configurable process: Variant-Rich BPMN language

There are many approaches to represent the variability of BPMN process [7] [9] [10]. All these approaches are derived from the Variant Rich BPMN (VR-BPMN) language [5]. The VR-BPMN extends BPMN to support the variability of business processes using annotation technique. It allows representing three concepts of variability, a variation point (alternative or optional), a variant (default or simple variant) and the relationship between variation point and variants (encapsulation, extension, inheritance). It was used in several case studies within the automotive field [6] and E-healthcare [11]. All the variability representation stereotypes are described in the following table (cf. Table I):

TABLE I. VARIABILITY REPRESENTATION STEREOTYPES OF THE VR-BPMN

Configurable elements		***Variability representation Stereotypes***
Variation point activity	**Alternative**	-« VarPoint» defines an alternative variation point activity -« Abstract» defines an abstract variation point activity with several implementations.
	Optional	-«Optional »- defines an optional variation point activity that can be extended by several variants.
Variant		-«Default » represents the default realization of a variation point activity. -«Variant » represents the realization of a variation point activity
Association {variation point variant}		-« implementation » is used to associate a variant activity with an abstract variation point activity. -« inheritance » is used to signify that a variant activity is a type of a variation point activity -« extension » is used to associate a variant activity with an optional variation point activity.

For each variant activity, a feature is associated to define the selection condition of the alternative activity. In this work, the Variant-Rich BPMN is used to represent the variability of all perspectives of business processes (functional, behavioral, organizational and informational) [13], unlike the generation approaches that are limited to activity and data transformations. A complete representation of configurable process will allow us to treat the generation of services in a broader sense.

3) Representing configurable services: VarSOAML language

Recently, with the emergence of reuse in SOA, some approaches, yet few, became interested in modeling services supporting variability. The approaches that propose an extension of SoaML language to support the variability of services are thus examined. The VarSOAML language [25] is adopted. It represents the variability of all SoaML service elements, to cover four views of service, namely the business view, structural, functional and composition. To extend the elements of SoaML, VarSOAML uses UML stereotypes. Different representation stereotypes of service elements (contract participant, message, service and operation interface) are described below (cf. Table II).

TABLE II. REPRESENTATION STEREOTYPES OF VARSOAML

Representation stereotypes of variable service elements	Definition
« VariableContract «	Describes the variability of collaboration.
« VariableInterface »	Define a variable service interface
« VariationOperation »	Defines a variation required or consumer operation
« VariableMessage »	Defines a variable message type of service.
« VariantOperation »	Defines a variant operation.
« VariationType »	Defines a variation data.
« VariantType »	Define a variant data.
« VariableParticipant»	Represents a variable participant of service

It appears that the VARSOAML language is the richest language in terms of representation, because it covers four views of service, namely the Service View (contract variable), functional view (variable interface, variation / variant operation, variation / variant type), structural view (variable participant) and the composition view (UML activity diagram).

4) Generation approach: MDA

MDA can be defined as the achievement of the MDD approach (Model Driven Developement) around a set of OMG standards such as MOF (Model Object Facility), UML, XMI and OCL enabling a new model based development approach.MDA defines three types of models [24]:

- CIM (Computation Independent Model): It represents the business requirements of a system. This is a model of business requirements defining the business interactions and business tasks of a system, without describing its structure or its implementation. In object-oriented approaches, CIM is represented by the use case diagram, while in service-oriented approaches CIM is represented by BPMN business process models [14] [17] or UML activity diagram [25].
- PIM (Platform Independent Model): it represents a model describing the business logic of a system, independently of any technology. It allows describing the structure of the entities which constitute the system. In object-oriented approaches, the UML class diagram is often used at this level, while in the service-oriented approach; the PIM is represented by SoaML models [14].
- PSM (Platform Specific Model) is a model that represents an implementation of a system according to a particular technology. MDA offers UML profiles to create these models, such as EJB profile (Enterprise Java Beans).

In order to establish traceability between CIM, PIM and PSM levels, MDA proposes the model transformation concept. These models must conform to their meta models. Thus, a metamodel is a model of a modeling language. The model transformation can be of three types:

- Simple transformation (1 to 1): it combines every element of source model with at most one element of the target model. An example of this transformation is the transformation of a UML class in a Java class.
- Multiple transformations (M to N): it takes as input a set of elements of the source model and produces a set of elements of the target model. Sometimes this transformation can be a type of composing models (1 to N) or merging models (N to 1).
- Update transformation: it is dedicated for changing a model by adding, modifying or deleting some of its elements.

In this paper, the composing models category (1 to N) is adopted. In fact, the idea is to transform a VR-BPMN configurable process model to four VarSOAML models which represents configurable services. The paper aims to offer a service generation method to generate configurable services from a configurable process, covering all perspectives of a configurable process. Thus, it is important to mention that the proposed approach will allow generating all the models representing a configurable service including contract, interface, Message Type and participants. These models will be transformed into configurable web services.

III. STATE OF THE ART

In this section, existing solutions in service generation are analyzed, and are evaluated how suitable for the purposes of this work they are. This analysis also provides valuable input regarding the requirements of this proposal.

All existing works on service generation [14] [16] [17] [18] [25] adopt MDA approach. Before analyzing these approaches, the evaluation criteria are listed below:

- **Transformation level** is about two kinds:
  - From business process model to service models (CIM2PIM)
  - From Services models to web services (PIM2PSM).
- **Representation language** determines the language used.
- **Variability** specifies whether the management of business process and service variability is assured.
- **Perspective** mentions the elements supported by the mapping rules.
- **Method** indicates if the approach proposes a method to assist the designers when generating services.

The following existing works are presented:

- MINERVA Framework [14]

In this work, authors develop a Framework called MINERVA (Model Driven & Service Oriented Framework for the continuous Business Process Improvement & related tools). MINERVA generates, from a BPMN business process models, SoaML service models (corresponding to CIMtoPIM transformation). The SoaML models are then transformed to

execution models represented in WSBPEL or XPDL (corresponding to PIMtoPSM transformation).The authors use a combination of Eclipse plugins (BPMN modeling, Medini QVT, Magic Draw and Model Pro) to implement their solution. However, no explicit definition of mapping rules at different levels of the Framework was found. Moreover, the approach does not support the concept of variability.

- BPMN-SoaML mapping [21]

In this work, authors define mapping rules for transforming BPMN models toSoaML models (CIMtoPIM).Thus, authors focus on the mapping of activities, Pool, Message Flow, and ignore the data and sub-processes. Regarding the target model, they focus on the contract, the service interface, the participant and the service architecture, messages and the choreographies. The rules are implemented using ATL (Atlas Transformation Language). However, the approach does not support the concept of variability.

-Business –IT alignment [17]

This approach is part of the business IT alignment based on a BPM-SOA convergence. In this work, authors propose a method to implement the business process as a service using the MDA approach. This work covers the mapping of the main perspectives of BPMN process models (business, sub-processes, Pool, Lan and process fragment) to the service model elements (Service Architecture, interface, contract and participant). A detailed definition of mapping rules is provided, as well as implementation in ATL language. However, this approach does not support the concept of variability.

- BPMN-SCA (Service Component Architecture) [16]

This work provides a mapping between business process models represented by BPMN and service models represented by SCA, which provides composition of applications using the principles of SOA. This approach focuses primarily on collaborative elements, participant and activity. The functional view and the service view is not supported by the approach. In addition, BPMN models supporting variability are not included.

- SVDEV [25]

This work proposes a development method SVDev (Service Variability Development) using the MDA approach. The SVDev development method is organized according to the levels of MDA:

- CIM level: it focuses on the study of the preliminary analysis and the specification of the business process models and classes.
- PIM level: it represents services (supporting variability) by VarSOAml. These models are organized in terms of views (functional, service, structure and composition).
- PSM level: it implements services witch support the variability by using VarWebService.

The Table III provides a summary of the following comparative study:

TABLE III. THE SERVICE GENERATION WORKS COMPARATIVE STUDY

Approaches	Transformation levels		Representation language		Variability		Perspective		Method
	CIM 2PIM	PIM2PSM	BP	Service	BP	Service	BP	Service	
[14]	Yes	Yes	BPMN	SOAml	-	-	-	-	No
[21]	Yes	-	BPMN	SOAml	-	-	Activity, Pool and message	contract, interface participant, service architecture, messages and choregraphies.	No
[17]	Yes	Yes	BPMN	SOAml	-	-	activité, sub processes, Pool, Lan	Service architecture, interface, contract and participant	Yes
[16]	Yes	-	BPMN	SCA	-	-	Activity, collaboration , conversation and participant	Component, and services	No
[25]	-	Yes	DA UML	VARSOAml	Yes	Yes	-	Participant variable, contrat variable, interface de service variable et message type variable	Yes

All approaches cover CIM2PIM level except SVDEV which covers only PIM2PSM. The most approaches use BPMN for business process models and SoaML for service models. Only SVDEV uses VARSOAML for service models. None of the approaches uses BPMN with supporting variability. Only the work [17] supports the mapping of all elements.

None of the analyzed works was found suitable to fulfill the needs of business IT alignment with supporting variability. In order to overcome these limitations, an MDA based method for aligning configurable services to configurable processes is proposed. This approach should enable the generation of configurable services from a configurable process while covering the generation of all views of a service.

IV. MDA Based Method for Configurable Services Generation

The main contribution that we look forward to in this work is the MDA based generation method (cf. Fig. 1) which is developed for PAIS designers. In what follows, all method steps are given.

Thus, the proposed method consists of the following steps:

- (1) Configurable process modeling
- (2) Decomposition of the configurable process into several fragments, each fragment represents a configurable service
- (3) Configurable service generation which consists of two sub steps :
  - VarSOAML configurable service generation which describes the configurable services associated with the identified fragments.
  - Configurable web services generation which corresponds to the service implementation.

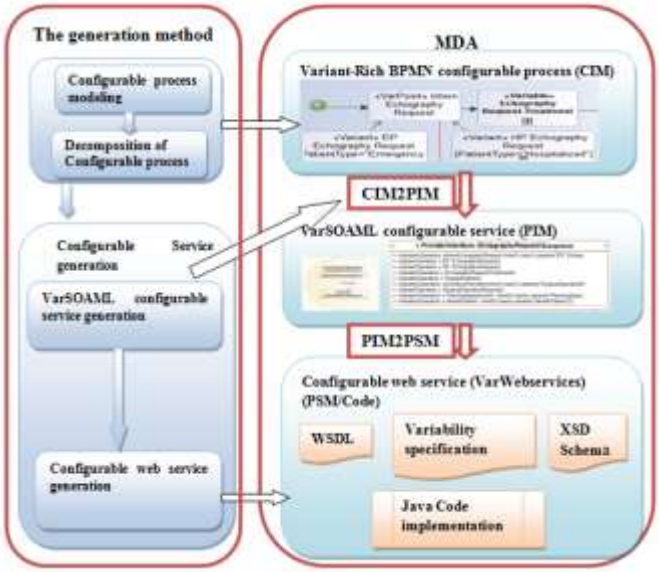


Fig. 1. The proposed MDA based method

A. Configurable process modeling

In order to illustrate the various steps of generation of configurable services, an echography request process represented in Variant-Rich BPMN (see Fig. 2) is used. This configurable process includes three variants: echography request of a simple patient (variant 1), echography request of a hospitalized patient (variant 2) and echography request of a emergency patient (variant 3):

- Variant 1: the request is made by the patient. It is then received by the assistant who is responsible for the management of patients and then studied by a hospital actor (the Cardiologist Doctor) who is responsible of patient examination.
- Variant 2: the request is made by a hospital Actor (doctor) and received by the assistant. The same process as a simple patient is applied.
- Variant 3: the request is sent by a hospital actor (emergency doctor), then it is sent directly to the cardiologist to perform an emergency examination.

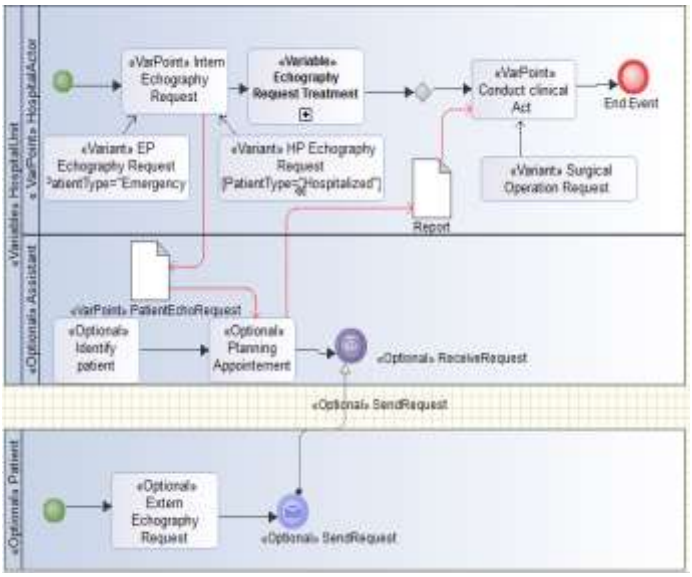


Fig. 2. An echography request configurable process

For each variation point activity, the type of the variation point (through the stereotype) and the associated variants are indicated, and for each variant activity, the feature helping the designer in the variability resolution step ({PatientType = "Hospitalized patient"}) is represented.

B. Decomposition of configurable process

The decomposition of a business process is to extract from a business process model, a set of fragments that encapsulate a business objective. A fragment is identified from a series of sequential activities or from a Gateway. A fragment corresponds to a business service [26].

In this work, the decomposition of a configurable process (shown in Variant-Rich BPMN) is to extract several configurable fragments (supporting variability). These configurable fragments correspond to configurable composite activities. A composite activity is called configurable if it contains at least one variation point or variable activity. A configurable fragment is identified from a series of sequential activities carried out by the same participant. For each configurable fragment identified, a configurable service will be associated.

Thus, the configurable process can be decomposed as follows (see Fig. 3):

- The composite activity

« EchographyRequestManagement» consists of:

- A variation point activity «InternEchographyRequest»
- A Variable activity «EchographyRequestTreatment»
- A variation point activity «Conduct Clinical Act»
- An optional activity «Planning Appointment»
- An optional activity «IdentifyPatient»

- The composite activity «PatientEchographyPatient» consists of:

- An optional activity "ExternEchographyRequest»

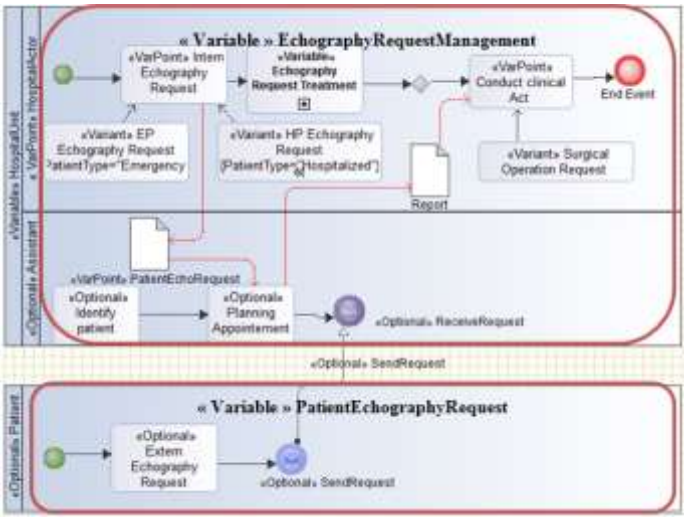


Fig. 3. Decomposition of the configurable process

Each configurable composite activity is associated with a configurable service. Two configurable services:

- EchographyRequestManagement
- PatientEchographyRequest

For each identified configurable composite activity, the service interface, the messages, the service contract and participants are generated.

C. VarSOAML configurable service generation

Before generating a VarSOAML models of a configurable service, it is first to identify the elements of the source metamodel and the target metamodel of those affected by this generation.

The type of transformation which is operated is the multiple transformation (1 to 3).A Variant-Rich BPMN model is transformed to four VarSOAML models by applying generation rules.

Example of the proposed generation rule [15]:

The rule for the transformation of the VariableCompositeActivity element to ProviderInterface or ConsumerInterface elements is given, as well as VariableCompositeActivity element is shown in VarSOAML by two interfaces: Provider Interface and Consumer Interface.

```
Rule name:
VariableCompositeActivity2ConsumerInterface
&ProviderInterface
Input element: VariableCompositeActivity
Output element: ConsumerInterface or
ProviderInterface
        For each
        VariableCompositeActivityelement Do
           Create an element of
           ConsumerInterfaceOrProviderInterface
           types
           The name of the ConsumerInterfaceOr
           Provider element is the name of the
           VariableCompositeActivityelement
           If (the IncomingMattribute is Null)
           of the first SimpleActivity or
           VariationPointActivity or Event
           elements contained in the
           VariableCompositeActivityThen
            Create ConsumerInterface element
           Else
           Create ProviderInterface element
           Apply
           VariableCompositeActivity2VariableIn
           terface//Create aVariableInterface
           element which represents the service
           interfacewhich provides the
           ProviderInterface
           End If
        For each element SimpleActivity element
        in VariableCompositeActivityDo
           Apply SimpleActivity2Operation
        End For
        For each VariationPointActivity element
        Do
           Apply
           VariationPointActivity2VariationOper
           ation
        End For
```

The example of VarSOAML models which represent the service associated with the composite activity «EchographyRequestManagement» is given bellow.

a) Service contract model

By applying the rule

VariableCompositeActivity2VariableContract [15] the service contract model «EchographyRequestManagement» can be generated (cf. Fig. 4).

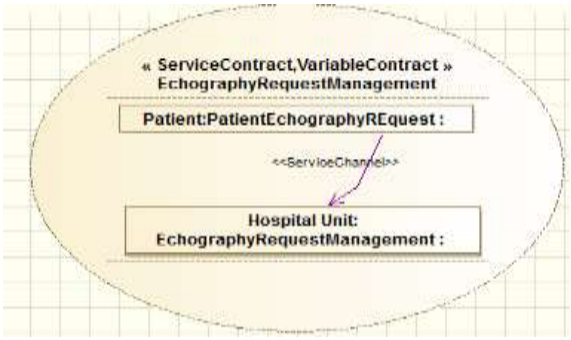


Fig. 4. Service contract model

b) Service Interface Model

By applying the rules proposed in [15]:

VariableCompositeActivity2VariableInterface

- VariableCompositeActivity2ProviderInterface & ConsumerInterface
- VariationPointSimpleActivity2VariationOperation
- VariantSimpleActivity2VariantOperation
- SimpleActivity2Operation

The configurable service interface is given bellow (cf. Fig. 5).

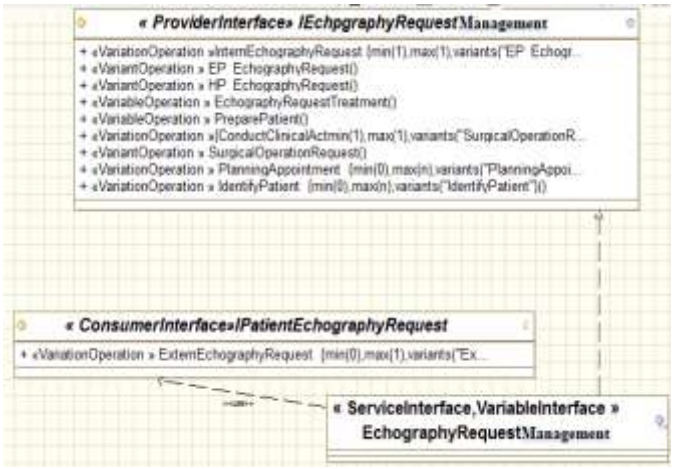


Fig. 5. Service Interface model

This interface uses the required interface «IPatientEchographyRequest» and provides «IEchographyRequestManagement».

c) Message Type Model

By applying the rules [15]:

- VariationPointActivity2VariableMessage
- VariableActivity2VariableMessage
- VariationPointDataObjectInput2VariationType
- VariationPointSimpleDataInput2VariableAttribut

The Message Type model is exposed in (cf. Fig. 6)

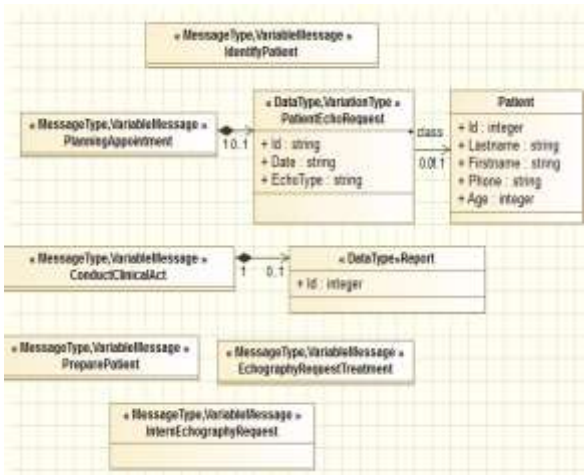


Fig. 6. Message Type model

The model describes the variables and simple messages. The model also describes the data contained in the message.

D. Configurable web services generation

In this section, the generation of configurable web services (called VarWebservice) from VarSOAML models is detailed.

The VarWebService generation approach proposed in [25] is applied.

This work establishes the transformation of VarSOAML models to configurable web services. Thus, a web service is defined as a service that is accessible via the Internet and uses the XML standard. It is a software module that exposes the interface through a WSDL (Web Service Description Language). WSDL is a language for describing all operations and messages that can be exchanged [27].

The generation of configurable web services requires generation of the following files:

- WSDL file
- Variability specification file associated with the WSDL file
- XSD schema which describes data which are used by the web service
- Java code implementation

1) WSDL file generation

The WSDL file (cf. Fig. 7) contains all the messages and operations. The following generation rules are applied:

- ServiceInterface2WSDL
- Operation2Operation
- MessageType2Message

Example of generation rule: ServiceInterface2WSDL

The ServiceInterface2WSDL rule enables the creation of instances of Binding, PortType, Schema Types ('WSDL') elements from the instance of ServiceInterface element.

```
Name rule: ServiceInterface2WSDL
Input Elements: ServiceInterface (VarSOAML)
Output elements: Binding, PortType, Schema
Types, Definition ('WSDL'
For each ServiceInterface Do
      Create an element of wsdl
            Create the Binding attribut
            The name of Binding is the name
            of ServiceInterface concatenated
            with the string 'Binding'
            Create the style attribute
            The name takes the value
            "document"
            For each Operation Do
                  Apply Operation2Operation
                  rule
            End For
End For
```

```
<?xml version="1.0" encoding="UTF-8"?>
<wsdl:definitions
name="EchographyRequestManagement"
targetNamespace="urn:// 
EchographyRequestManagement.wsdl"
xmlns:soap="http://schemas.xmlsoap.org/wsdl/so
ap/"
xmlns:xsd="http://www.w3.org/2001/XMLSchema"
xmlns:ps="urn://EchographyRequestManagement
Schema.xsd"
xmlns:tns="urn:// 
EchographyRequestManagement.wsdl"
xmlns:wsdl=http://schemas.xmlsoap.org/wsdl/
xmlns="http://schemas.xmlsoap.org/wsdl/">
<wsdl:import namespace="urn:// 
EchographyRequestManagement Schema.xsd"
location=" EchographyRequestManagement
Schema.xsd"> </wsdl:import>
<wsdl:types></wsdl:types>
<wsdl:message name="return">
<wsdl:part name="partreturn" type="xsd:int">
</wsdl:part>
<wsdl:message name="PlanningAppointment">
<wsdl:part name="part PlanningAppointment"
type="ps:PatientEchoRequest"/>
</wsdl:message>
<wsdl:message name="ConductClinicalAct">
<wsdl:part name= "part ConductClinicalAct"
type="ps: Report"/>
</wsdl:message>
<wsdl:portType
name="EchographyRequestManagementPortType">
<wsdl:operation name="PlanningAppointment">
<wsdl:input message="tns: PlanningAppointment"
name="PlanningAppointment_Request"/>
<wsdl:output message="tns:return"
name="PlanningAppointment_Response"/>
</wsdl:operation>
<wsdl:operation name="ConductClinicalAct">
<wsdl:input message="tns: ConductClinicalAct"
name="ConductClinicalAct_Request"/>
<wsdl:output message="tns:return" name="
ConductClinicalAct_Response"/>
</wsdl:operation>
</wsdl:portType> </wsdl:definitions> …
```

Fig. 7. An extract of the WSDL file

The VarSOAML source elements affected by this generation are: the interface provided by the service interface and messages associated with this interface.

2) Generation of the variability specification associated with the WSDL file

The specification of the variability associated with the WSDL file describes the variable operations, messages, types and variables attributes (cf. Fig. 8).

```
<variability service
="EchographyRequestManagement" name
="EchographyRequestManagementVariability">
<operations>
<variationOperation name =
"InternEchographyRequest" min = "1" max =
"1" portype =
"EchographyRequestManagementPortType">
</variationOperation>
<variantOperation name = "EP
EchographyRequest"
</variantOperation>
</operations>
<messages>
<variablemessage name = "
PlanningAppointment" scope ="configurable"
boundElement = " PlanningAppointment"
description ="">
<types> <type>RequestEcho</type> </types>
</variablemessage>
<variablemessage name = "
ConductClinicalAct" scope ="configurable"
boundElement = "ConductClinicalAct"
description ="">
<types> <type>Report</type>
</messages>
…
</variability>
```

Fig. 8. An extract of the Variability specification associated with the WSDL file

3) XSD schema generation

The XSD schema describes all data types used by a web service.

In order to generate the XSD file (cf. Fig. 9), the generation rules proposed in [25] are applied:

- ServiceInterface2Schema
- MessageType2Message
- Attribute2Element
- DataType2Type
- PType2SimpleType

Example of the generation rule: Data Type2Type:

The DataType2Type rule allows for creating an instance of the web service ComplexType from an instance of the VarSOAML DataType element.

```
Name rule: DataType2Type
Input Element: DataType (VarSOAML)
Output element: ComplexType ('XSD')
```

```
For each DataType element Do
       Create ComplexType element
       The name of the ComplexType is the name
       of the Datatype
       For each attribute element Do
              Apply the Attribute2Element rule
       End For
End For
```

```
<?xml version="1.0" encoding="UTF-8"?>
<xsd:schema
xmlns:xsd="http://www.w3.org/2001/XMLSchema"
elementFormDefault="qualified"
xmlns:tns="urn://
EchographyRequestManagement Schema.xsd"
targetNamespace="urn://
EchographyRequestManagementSchema.xsd">
<xsd:complexType
name="sequencePatientEchoRequest">
<xsd:sequence>
<xsd:element name="PatientEchoRequest"
type="tns: PatientEchoRequest" />
</xsd:sequence>
</xsd:complexType>
<xsd:complexType name="PatientEchoRequest">
<xsd:sequence>
<xsd:element name="id" type="xsd:integer" />
<xsd:element name="Date" type="xsd:string"
/>
<xsd:element name="EchoType"
type="xsd:String" />
<xsd:element name="Patient"
type="xsd:Patient"/>
</xsd:sequence>
</xsd:complexType>
<xsd:complexType name="Patient">
<xsd:sequence>
<xsd:element name="id" type="xsd:string" />
<xsd:element name="Lastname"
type="xsd:string" />
<xsd:element name="Firstname"
type="xsd:string"/>
<xsd:element name="Phone" type="xsd:string"
/>
<xsd:element name="Age" type="xsd:integer"
/>
</xsd:complexType>
<xsd:complexType name="Report">
<xsd:sequence>
<xsd:element name="id" type="xsd:string" />
<xsd:element name="ReportSubject"
type="xsd:string" />
<xsd:element name="Date" type="xsd:string"
/>
<xsd:element name="Act" type="xsd:string" />
</xsd:schema>
```

Fig. 9. An extract of the XSD schema associated with the WSDL file

4) Java code implementation of configurable web service

The following generation rules are applied in order to generate the java code implementation [25] (cf. Fig. 10):

- ServiceInterface2Interface
- Operation2WebMethod
- Attribut2WebParam

Example of the generation Rule: Operation2Method

The Operation2Method rule allows creating of the Method instance from Operation.

```
Name rule: Operation2Method
Input Element: Operation (VarSOAML)
Output elements: Method, WebMethods ('JAXWS')
Description:
For each Operation element Do
       Create Method element
       The name of Method parameters is the
       name of the parameters and return type
       of the Operation element
       For each Parameters Do
              Apply Param2Par rule
                     Creates WebMethod
                     The name of OperationName
                     is the name of the
                     operation element
       End For

End For
```

The java code of the web service EchographyRequestManagement is generated (cf. Fig. 10):

```
package EchoRequestWebService;
import javax.jws.*;
import javax.jws.soap.SOAPBinding;
import javax.xml.ws.Endpoint;
@WebService(serviceName="EchographyRequestMan
agementService")
public class EchographyRequestManagementClass
implements
EchographyRequestManagementInterface {
@VariantOperation
(boundOperation="EP_EchographyRequestPatient"
)
public EP_EchographyRequestPatient()
{ …
}
@VariantOperation
(boundOperation="HP_EchographyRequestPatient"
)
public HP_EchographyRequestPatien()
{ …
}
}
```

Fig. 10. An extract of the Java Code implementation

The advantage of this method is that it covers all stages of configurable services development, from configurable process modeling to configurable web services implementation which is described by the WSDL file, the variability specification, the XSD schema and the Java code implementation.

V. Conclusion

Many solutions for business IT alignment have been proposed. However, some limitations such as the weak support of business process and service variability are underlined. The MDA based method for service generation ensures the business IT alignment with managing variability.

It allows for decomposing configurable processes into set of configurable composite activities. Thus, a set of configurable services VarSOAML generation rules were proposed with the aim of generating different VarSOAML models. The advantage is that these models are then automatically implemented as configurable web services which requires the generation of the WSDL file, the variability specification, the XSD file and the java code implementation. In this sense, VarSOAML models conduct the implementation of configurable web services.

Indeed, the generation of configurable services from a configurable process provides better synchronization of changes between the business and IT layers, it will facilitate the propagation of changes from configurable process to configurable services.

Furthermore, the proposed business IT alignment is not sufficient as it excludes changes that may come from the service layer. As future work, it will be necessary to offer a bottom up approach to align business process with services. Another possible improvement is to incorporate the semantic aspect to enrich configurable services in the context of business IT alignment. This can allow for developing intelligent configurable web services.

REFERENCE

[1] M. Dumas, W.M.P. Van der Aalst and A.H.M. ter Hofstede,"Process-Aware Information Systems", Wiley, 2005.

[2] C. Ayora, V. Torres, B. Weber, M. Reichert ,V. Pelechano, "Enhancing Modeling and Change Patterns. BMDS/EMMSAD", CAiSE 2013, Valencia, Spain, pp.246-260, 17-18 Juin, 2013.

[3] M. La Rosa,"Managing variability in PAIS", PHD thesis, Faculty of Science and Technology, Queensland University of Technology, Brisbane, Australia, 25 Mars 2009.

[4] M. Rosemann , W. M.P Van der Aalst,"A Configurable Reference Modelling Language ", Information Systems Vol.32, N°1, pp.1–23, 2007.

[5] A. Schnieders , F. Puhlmann,"Variability modeling and product derivation in ebusiness process families", Technologies Business Information System Book, Springer, pp. 63–74, 2007.

[6] A. Hallerbach ,T. Bauer, M. Reichert,"Capturing variability in business process models: The Provop approach", Journal of Software Maintenance and Evolution: Research and Practice, Vol.22, N°7, pp.519–546, 2010.

[7] V. Kulkarni, S. barat,"Business process families using model-driven techniques Lecture Notes in Business Information Processing", Vol.66, pp 314-325, 2011.

[8] A. Kumar, W. Yao,"Design and management of flexible process variants using templates and rules,», Computers in Industry Vol.63, N°2, pp.112–130, 2012.

[9] T. Nguyen , A. Colman, J. Han ,"Comprehensive Variability Modeling and Management for Customizable Process-Based Service Compositions", in : A. Bouguettaya, Q. Z. Sheng, F. Daniel, Web Services Foundations, Springer, pp.507–534, 2014.

[10] A. Yousfi, R. Saidi and A.K. Dey,"Variability patterns for business processes in BPMN. *Information Systems and e-Business Management*, pp:1-25,2015.

[11] C. Ayora, V. Torres, J L. De la Vara, V. Pelechano,"Variability management in process families through change patterns", Information and Software Technology, Vol.74, pp. 86-104, 30 June 2016.

[12] B. Weber, M. Reichert, S. Rinderle," Change Patterns and Change Support Features: Enhancing Flexibility in Process-Aware Information Systems", Data and Knowledge Engineering, Vol.66, N°3, pp. 438–466, 2008.

[13] H. Sbai, M. Fredj, L. Kjiri," A pattern based methodology for evolution management in business process reuse", IJCSI International Journal of Computer Science Issues, Vol. 11, Issue 1, N°1, pp. 211-220, January 2014.

[14] A. Delgado, F. Ruiz, I.G.R. de Guzman, M. Piattini,"A model-driven and service-oriented framework for the business process improvement", Journal of Systems Integration, Vol.1, No° 3, pp. 45–55, 2010.

[15] H. Sbai, M. Fredj, B. Chakir,"Generating services supporting variability from configurable process model", Journal of Theoretical and Applied Information Technology, Vol. 72, N°2, pp. 111-124, February 2015.

[16] K. Da man, F. Charoy, C. Godart,"Alignment and change propagation between business processes and service-oriented architectures», International Conference on Service Computing (SCC'13), Santa Clara, CA, United States, pp. 168–175, 27 June - 02 July, 2013.

[17] B. Elvesæter, D. Panfilenko , S. Jacobi and C. Hahn,"Aligning business and IT models in service-oriented architectures using BPMN and SoaML,"Proceedings of the First International Workshop on Model-Driven Interoperability (MDI '10), Oslo, Norway, 3-5 Octobre, 2010.

[18] A. Kabzeva and P. M¨uller," Toward Generic Dependency Management for Evolution Support of Inter-Domain Service-Oriented Applications", European Conference on Service-Oriented and Cloud Computing (ESOCC 2012), Bertinoro , Italy, pp.35-40 ,2012.

[19] Y. Wang, J. Yang, W. Zhao,"Change impact analysi"s for service based business processes. In the IEEE International Conference on Service-Oriented Computing and Applications (SOCA), pp. 1-8, 13 December, 2010.

[20] J. Simonin, P. Picouet, J.M. Jézéquel, "Conception fonctionnelle de services d'entreprise fondée sur l'alignement entre cœur de métier et système d'information", Ingénierie des systèmes d'information, Vol. 15, N°4, pp.37-61, 2010.

[21] Y. Lemrabet, J. Touzi, D. Clin, M. Bigand, J.P. Bourey, "Mapping of bpmn models into uml models using soaml profile", 8th International Conference of Modeling and Simulation (MOSIM'10), Hammamet, Tunisia, pp. 10-12 Mai, 2010.

[22] BPMN, Business Process Modeling and Notation, Version 2.0. Object Management Group (OMG) http ://www.bpmn.org/ (dernière consultation Mars 2015).

[23] SOAML, "Service oriented architecture modeling language (SoaML) – specification for the uml profile and metamodel for services (UPMS)", Version 1.0, 2012.

[24] J. Bézivin, "Sur les principes de base de l'ingénierie des modèles", ISSN 1262-1137, Vol. 10, No 4, pp. 145-156, 2004.

[25] B. Chakir, "Contribution à l'amélioration de la variabilité des services par la gestion de la variabilité ", doctoral dissertation, Mohammed V University of Rabat, March 2014.

[26] M. Radgui, "Décomposition et adaptation de processus métiers BPMN pour des systèmes d'information flexibles ", Mohammed V university of Rabat, Morocco, June 2015.

[27] W3C, Web Services Definition Language (WSDL) 1.2, : http://www.w3.org/TR/2003/WD-wsdl12-20030611/, 2003.

Performance Metrics for Decision Support in Big Data vs. Traditional RDBMS Tools & Technologies

Alazar Baharu
Lecturer, Computer Science & IT Department,
AMiT, AMU
Ethiopia

Durga Prasad Sharma
Professor, CS&IT
AMiT, AMU
MOEFDRE

***Abstract*—In IT industry research communities and data scientists have observed that Big Data has challenged the legacy of solutions. 'Big Data' term used for any collection of data or data sets which is so large and complex and difficult to process and manage using traditional data processing applications and existing Relational Data Base Management Systems (RDBMSs). In Big Data; the most important challenges include analysis, capture, curation, search, sharing, storage, transfer, visualization and privacy. As the data increases in various dimensions with various features like structured, semi structured and unstructured with high velocity, high volume and high variety; the RDBMSs face another fold of challenges to be studied and analyzed. Due to the aforesaid limitations of RDBMSs, data scientists and information managers forced to rethink about alternative solutions for handling such data with 3Vs.Initially research study focused on to develop an intelligent base for decision makers so that alternative solutions for long term suitable solutions and handle the data and information with 3Vs can be designed. In this research attempts has been made to analyze the feature based capabilities of RDBMSs and then performance experimentation, observation and analysis has been done with Big Data handling tools and technologies. The features considered for scientific observation and analysis were resource consumption, execution time, on demand scalability, maximum data size, structure of the data, data visualization, and ease of deployment, cost and security. Finally the research provides a decision support metrics for decision makers in selecting the appropriate tool or technology based on the nature of data to be handled in the target organizations.**

***Keywords*—Big Data; RDBMSs; big data tools; Variety; velocity; volume; Metrics**

I. Introduction

Currently big data analysis is an emerging domain of research and has become a new paradigm for business intelligence, predictions and forecasting in salient disciplines. In order to choose appropriate technology for data capture, curation and analysis; still there no clear decision support metrics to assist top level executives. A strong need is anticipated for development of a decision support system metrics for organizations who wants to handle different size data sets with different types and varied velocity and volume. There is a strong need for research oriented data handling mechanisms in order to support top level technocrats and executives for their decision making processes.

In the end of 1959 scientists have tried to trouble shoot these problems related to huge amount of data handling by emerging hierarchical DBMS's within organization having large amount of data with high computing power. This phenomenon was continued from 1960's – 1970's. After that a new era emerged i.e. was the era of EF Cod's RDBMSs that overcome the limitations of DBMSs [1]**.** These solutions worked fine tuned till 2007 but after that there was again a limitation i.e. how to handle a huge amount of data which is un-structured or semi structured and has been increasing in Zettabyte per day with critical time complexity due to the introduction of different technologies such as e-commerce, smart city surveillance camera, GPS systems and social networking Medias. The world's technological per-capita capacity to store information has roughly doubled every 40 months since the 1980s as of 2012, every day 2.5 Exabyte (2.5×10^{18}) of data were created; as of 2014, every day 2.3 Zettabyte (2.3×10^{21}) of data were created [2]**.**

This dramatic increase in data with high variety, high velocity and high volume became difficult to handle by existing principles and mechanisms therefore a concept of Big Data has been evolved.

When the scope and importance of this research is narrowed down in Ethiopian context, it has been observed that business houses, millionaires, decision makers here believe on numbers rather than theoretical aspects and predictions. However the thorough analysis of literature review clearly indicate that there is no or very limited research studies conducted in the area especially for Ethiopian need and context, which may be a serious issue in future endeavors and key indicator for Ethiopian investors, business houses and industrialists.

This research study strive to conduct the performance analysis of Big Data vs. RDBMS tools and technologies to develop a crystal clear performance metrics that can support the decision makers to select the appropriate tool or technology from amongst the RDBMS and Big Data. Further, the parameters considered in this research are time complexity of search queries, memory management, data visualizations, scalability, deployment cost etc.

II. Review of Literature

In the simplest way 58,300 results were found when searching for the term "difference between RDBMSs and Big Data" in Google. This can show as how much this topic is confusing and it needs to be clearly and scientifically explained [3]. In addition to this fact can also show as there is a gap of

concept and research works. This in turn confirms the research ability of the topic selected. Research related to Big Data emerged in the 1970s but has seen an explosion of publications since 2008. Big Data is an all-encompassing term for any collection of data sets so large or complex that it becomes difficult to process using traditional data processing applications like RDBMSs and Data warehousing tools and technologies. The challenges include analysis, capture, curation, search, sharing, storage, transfer, visualization, and privacy violations. If this flood of Big Data challenges the legacy of the RDBMSs; a solution is needed for long term sustainability in order to gain the full potentials of hidden insights in Big Data. Data is exploding so fast and the promise of deeper insights is so compelling that IT managers are highly motivated to turn big data into an asset they can manage and exploit for their organizations. Emerging technologies such as the Hadoop framework and Map Reduce offer new and exciting ways to process and transform big data defied as complex, unstructured, or large amounts of data [4].Why can't an analyst utilize databases with lots of disks to do large-scale batch analysis? Why is Map Reduce needed? The answer to these questions comes from another trend in disk drives: seek time is improving more slowly than transfer rate. Seeking is the process of moving the disk's head to a particular place on the disk to read or write data. It characterizes the latency of a disk operation, whereas the transfer rate corresponds to a disk's bandwidth. In many ways, Map Reduce can be seen as a complement to an RDBMS. Map Reduce is a good fit for problems that need to analyze the whole dataset, in a batch fashion, particularly for ad-hoc analysis. An RDBMS is good for point queries or updates, where the dataset has been indexed to deliver low-latency retrieval and update times of a relatively small amount of Data. Relational data is often normalized to retain its integrity, and remove redundancy. Normalization poses problems for Map Reduce, since it makes reading a record a nonlocal operation, and one of the central assumptions that Map Reduce makes is that it is possible to perform (high-speed) streaming reads and writes [5].

Applying Big Data analytics to the fuel of development faces several challenges. Some relate to the data including its acquisition and sharing, and the overarching concern over privacy. Others pertain to its analysis [6].The most important affecting and challenging Factor in Big data is Privacy. It is the most sensitive issue, with conceptual, legal, and technological implications. However the three basic and important requirements for RDBMSs are confidentiality, integrity and availability. The stored data must be available when it is needed (availability), but only to authorized entities (confidentiality), and only modified by authorized entities (integrity)). Traditional relational database management systems (RDBMS), like Oracle, SQL and MySQL, have been well-developed to meet the three requirements. In addition, enterprise RDBMS are further required to have ACID properties, Atomic, Consistency, Isolation, and Durability, that guarantee that database transactions are processed reliably. With such desirable properties, RDBMS have been widely used as the dominant data storage choice). RDBMS now are facing major performance problems in processing exponential growth of unstructured data, such as documents, e-mail, multi-media or social media. Thus a new breed of non-relational, cloud-based distributed databases, called NoSQL, has emerged to satisfy the unprecedented needs for scalability, performance and storage)[7].

III. Research Methodology

This research study uses both qualitative and quantitative research methods. Quantitative technique used to collect and convert data into numerical form so that statistical calculations can be applied to conclusions.

Qualitative research methods applied to qualitatively analyze the research question and that of the quantitative research method helped to analyze based on numbers.

A. Data Collection Methods

The data collection methods used was formal interview with different professionals. Web articles, scholarly paper, white papers, flayers, company product specifications, and books related to the research topic were also analyzed critically as secondary sources of facts. In addition to this 60 IT professionals were selected randomly to gather facts about the commonly used Big Data and RDBMSs tools and technologies. The data set for experiment purpose is the *Olympic Games winners' dataset* which is acquired from the publically available data source *Talned.* Data analytical tools were selected based on the parameters like 1) Popularity statistics, 2) Market coverage statistics, 3) Commonly preferred Practices, 4) Professional recommendations statistics, 5) Professional recommendation collected through formal interview in the fact finding technique.

B. Organizing Data set

After collection of data sets; the data set has been organized in a manner to make it suitable for the experimentation purpose.

IV. Performing Experiment on the Dataset

Experiments & analysis on data sets were done both quantitatively and qualitatively. The parameters selected for comparison metrics were CPU consumption, Memory (RAM) consumption, Execution Time, Scalability, Inter-Operability, Ease of deployment, System Security, data visualization, maximum data size and Cost.

A. Experimental procedure

Among the selected parameters for the performance analysis of Big Data and RDBMSs tools& technologies; CPU Time, Private working set, Execution time of each tools were measured and recorded by using the windows task manager and windows performance monitoring tools. In this performance analysis the two operations i.e. WRITE and READ operations were selected. Finally the performance of each tool has been measured by firing Reading and Writing Queries, then the result has been recoded to compare the performance of each tools.

1) Qualitative analyses

For qualitative comparison metrics performance analysis, System security, ease of deployment, Inter-operability and data visualization support parameters were analyzed. In doing this massive review of literature and fact finding techniques have been used.

2) Quantitative analysis

For quantities comparison metrics& performance analysis, the measurement of execution time, memory consumption, CPU consumption, cost, Max size of data, Scalability and data visualization (both quantitatively and qualitatively) were analyzed.

3) Comparative Analysis &Drawing Decision Support Metrics (System)

After extensive review of related research contributions, detailed experimentation, performance measurement and analysis; a comparative analysis was done to draw a decision support metrics for the future decision makers to select the most appropriate and most feasible RDBMS or Big Data tool or technology; based on their organizational needs and selection parameters.

V. Selection, Experimentation and Discussion

A. Selection of Data analytical tools

The selection of the data analytics tools has been performed based on the two 1) Database engine ranking website and 2) From the formal interview conducted on selected IT professionals and data experts. According to the database engine ranking; website ranking of the database systems was done based on- 1) Number of mentions of the system on websites, 2) General interest in the system, 3) Frequency of technical discussions about the system, 4) Number of job offers, in which the system is mentioned and 5) Number of profiles in professional networks, in which the system is mentioned. Based on such criteria the top three RDBMSs found were following-

TABLE I. RDBMS Rankings According to the Database Ranking Web Site [Till 1/8/2015]

Rank	Last Month	DBMS	Database Model	Score	Changes
1.	1.	Oracle	Relational DBMS	1439.16	-20.63
2.	2.	MySQL	Relational DBMS	1277.51	+8.93
3.	3.	Microsoft SQL Server	Relational DBMS	1198.61	-1.44

In addition to the above scientific observation and further verification; the IT professionals were interviewed. Based on the collected feedback from the interview it was confirmed that the most popular types of RDBMSs were the once ranked by the DB ranking engine and 69 percent of respondents responded that the introduction of Big Data Analytical technologies will bring additional features to data science.

On the collected opinion data and its analysis the Oracle, MYSQL and MS SQL Server were selected as top three RDBMSs from ten RDBMS. Expert & user opinion analysis clearly indicates that domain specific people (experts and users) also like the RDBMSs in the same ranking manner i.e. first Oracle, second MySQL and third Microsoft SQL Server.

B. Selection of Big Data analytical tools

In computer and IT world there are several methods, techniques, tools and technologies which are used for database creation, storage, management and analysis of different types of data, information, text and documents to be analyzed. When the data is being generated with high volume, high variety and high velocity then alternative tools and technologies are available in the market for different kinds of analytics in real time manner. According to the Apache technology specifications; there are a number of different flavors and distributions of apache Hadoop that are available for Big Data analytics. Some of them include Amazon Web Services, Apache Bigtop, Cascading, Cloudera, Cloudspace, Datameer, Data Mine Lab, Datasalt, Hortonworks, HStreaming, IBM, MapR Technologies, Think Big Analytics, and WANdisco[8].

Among all these Hadoop flavors; Hortonworks data platform was selected as a most significant data analytics tool in this research study. The parameters used to select the Hortonworks were, ease of accessibility (i.e. open source), easy to deploy as it has user friendly and GUI interfaces, unstructured registration for accessibility, well established development community with sufficient deployment records in the real world and incorporation of free online and offline embedded tutorials. Rest of the big data analytical tools is very difficult to deploy them and to get them work on the machine specification prepared for the research study. So the selection of the data analytical tools from both parties looks like unbalanced but it was observed that each Big Data analytical tools uses the same technologies as a foundation like Apache Hadoop, Hive, PIG script, MAPR, Hcatlaog and so on[9].

C. Selection of supportive analytical tools

- *Navicat premium*

Among third party DBMS Connection and management tools; Navicat premium was selected to help in organizing the data. Navicat Premium is a database administration tool with 100,000 users across 7 continents in more than 138 countries and it allows to simultaneously connecting to MySQL, MariaDB, SQL Server, Oracle, PostgreSQL and SQLite databases from a single application [10].

- *Oracle Virtual machine*

There are a variety of virtualization tools available on the market among them Oracle virtual machine have been selected and used as a virtualization tool for hosting the Horton Works Sandbox [11].

- *Windows task manger*

Windows Task Manager has been used to display the programs, processes, and services that currently run on a computer. In addition to this Task Manager is used to monitor computer's performance. In monitoring resource usage of a given process, task manger used different metrics like CPU Usage, CPU Time, Memory - Working Set, Memory - Peak Working Set, Process time etc. This research used four task

manager columns like CPU Usage, CPU Time, Threads and Memory-Private Working Set.

- *Windows performance monitoring*

Windows Performance Monitor tool can be used to examine how programs running in a system are affecting the performance of the computer, both in real time and by collecting log data for later analysis. Windows Performance Monitor uses performance counters, event trace data, and configuration information, which can be combined into Data Collector Sets.

In this research study windows performance monitoring tools has been used to measure how much resources are used by a given process and to conform the results observed from the Task Manager [12].

D. Comparative and performance analysis

In this section, the RDBMS and Big data tools were compared to analyze the performance; based on selected. During analysis the measurement metrics; used are CPU USAGE, Memory (RAM) usage, execution time and number of threads of Big Data and RDBMSs tools and technologies.

1) A comparative analysis based on Structure of data

The structure of data which is generated by different sources was categorized as structured, semi structured and unstructured. In this study, it was assumed that if someone wants to analyze the data for getting further insights and knowledge discovery, it has to deal with these three structures of data. For that data analysis RDBMS and Big Data Analytical tools and technologies may play a greater role and contribution in the process of knowledge discovery and decision making. Till today, each tool has its own support related to the structure of the data and most of the RDBMS tools handle only structured data and they don't have any provision to analyze the semi-structured and unstructured data. However Big Data analytical tools and technologies supported all structures of data.

2) A comparative analysis on support of maximum data size

The data size limitation for a row and a column for each tool have been observed in the following table 2:

TABLE II. MAXIMUM DATA SIZE CAPABILITY

Tools Name	Max DB size	Max table size	Max row size	Max columns per row
MSSQL	524,272 TB	524,272 TB	8,060 bytes	30,000
MySQL	Unlimited	64 -256 TB	64 KB	4,096
Oracle	4 GB * block size/table size	4 GB * block size	8 KB	1,000
Horton works	Unlimited	Unlimited	Unlimited	Unlimited

Above table clearly indicates that BIG DATA technologies have created a pace for unlimited data representation and handling capabilities where the RDBMSs features have been comprises their limits in terms of data size.

3) A comparative analysis of CPU consumption, RAM consumption (PWS), Execution time and number of threads used

In any computation or communication, the quality of an algorithm or a given instruction are measured based on the amount of resources it consumes. Resources like Processor and Memory are the most important things to be measured when to evaluate the performance of a given instruction.

In this research study, a simple query prepared for data insertion and reading was used as instruction to be executed on both RDBMSs and Big Data tools and technologies to measure performance differences in the tools related to the consumption of CPU, RAM, Number of threads and execution time (elapsed) were measured by using data ranging from 100 up to 5,000,000rows in both read and write operations.

In measuring the performance of target tools& technologies; the query was fired using each tool. Afterwards, the windows task manager and windows performance monitoring tools were monitored and the effect of the fired query on resources like CPU and RAM (memory) were observed and recorded. In addition to this; the execution time and the no of threads used were also observed and recorded. The execution time for the query execution was recorded from the target RDBMS and BIG DATA tools and the number of threads used were recorded from the Performance monitoring tool data collector set log file.

The two queries executed in each tools are:

Query for inserting Writing data: INSERT [Olympic Athletes] ([Athlete], [Age], [Country], [Year], [Closing Ceremony Date], [Sport], [Gold Medals], [Silver Medals], [Bronze Medals], [Total Medals]) VALUES (N' Michael Phelps', 23, N' United States', N'2008', CAST(0x00009B0200000000 AS Date Time), N' Swimming', 8, 0, 0, 8)

*Query for reading (selecting) data: Select * from Olympic Athletes limit 100-5000000;*

Based on the above process the result of the performance evaluation and analysis was illustrated in the following Table 3.The Operations were divided in to two categories- READING and WRITING.

TABLE III. PERFORMANCE ANALYSIS OF EACH TOOL IN READ OPERATION

Name of the tool	No of rows	CPU time	PWS(KB)	Thread	Time(S)
MSSQL	100	2.9	144864	6	0.085
	1000	3.1	144884	7	0.106
	5000	6.5	144864	27	0.162
	10000	6.6	144890	17	0.228
	100000	6	144898	12	1.422
	1000000	12	144900	3	13.950
	5000000	48	144929	8	66.592
MySQL	100	3	230200	3	0.032
	1000	4.62	250300	3	0.004
	5000	6.15	260100	3	0.015
	10000	6.36	270000	3	0.029
	100000	12.72	360500	3	0.366
	1000000	23.08	390500	3	4.820
	5000000	30.74	430500	3	27.635
Oracle	100	5.37	400800	3	0.002
	1000	5.41	400900	3	0.009
	5000	7.73	410100	3	0.44
	10000	9.29	410700	3	0.082
	100000	23.3	650100	3	1.108
	1000000	37.51	304800	3	9.669
	5000000	42.57	136333	3	50.269
Horton works Hadoop	100	5.9	870900	3	66
	1000	11.3	876856	3	83
	5000	15.5	878698	3	89
	10000	19.9	879652	3	90.325
	100000	22.6	889453	3	110
	1000000	29.54	889657	3	128
	5000000	40.89	889759	3	170

Generally in read operation the performance of RDBMS increases even if the number of the data size increase up to max level, however Big Data tools and technologies performance somehow decreases as size of data increases.

The performance measurement of WRITE operation has been recorded and illustrated in table 4.as follows:

TABLE IV. PERFORMANCE MEASUREMENT IN WRITE

Name of the tool	No of rows to write	CPU Time (%)	PWS(KB)	Threads	Elapsed time(S)
MS SQL Server	100	0.6	262840	2	0.103
	1000	2	192492	3	0.893
	5000	2.2	237792	3	4.736
	10000	3.7	287916	3	9.843
MySQL Server	100	1.8	43668	2	2.115
	1000	12.7	43684	2	22.046
	5000	19.9	44698	2	137.599
	10000	22.2	44800	2	258.581
Oracle	100	1.7	162764	2	0.189
	1000	11.7	204556	2	1.184
	5000	15.9	311968	2	11.094
	10000	36.7	396964	2	16.995
Horton works data platform	100	0.5	125000	3	4
	1000	0.9	125162	3	9
	5000	3.6	125173	3	11
	10000	5.9	125189	3	14

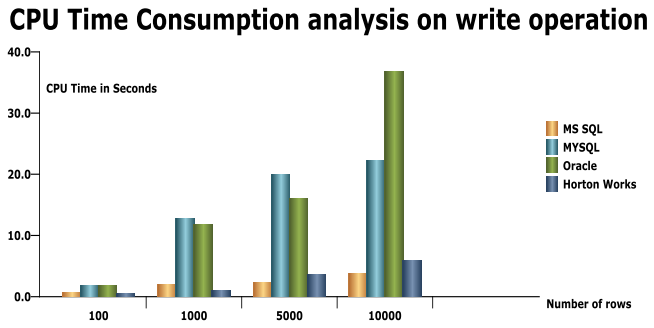


Fig. 1. CPU time consumption of each candidate tool in WRITE operation

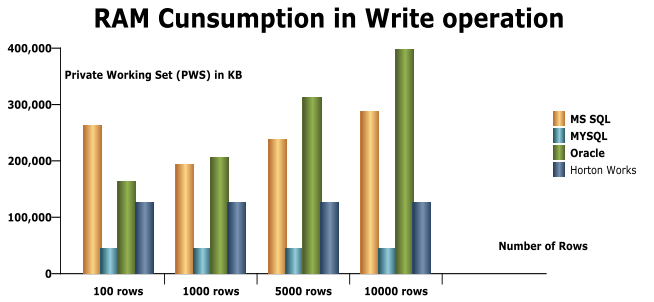


Fig. 2. RAM Consumption analysis of Each Candidate tool in WRITE operation

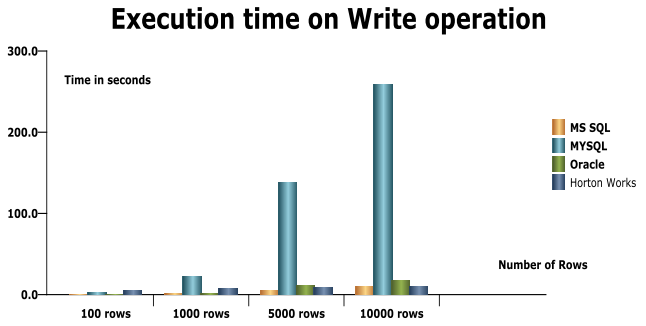


Fig. 3. Execution time analysis of each candidate tool in WRITE operation

In this phase it was observed that the performance of Big Data analytical tools and technologies increase in WRITE operation, however the performance of RDBMSs decrease when searching for the exact place or column table to put or to WRITE the data as the size of data increases.

E. A comparative analysis based on Software and operational cost

According to different scholar's ideas software were classified under two broad. These categories are open source and proprietary software.

The purchase cost of the selected data analytics tools has been summarized on the following table 5:

TABLE V. COST ANALYSIS OF EACH TOOLS AND TECHNOLOGIES

Name of the tools	License type	Purchase coast
MYSQL	Open Source	Free
MSSQL	Proprietary	$ 2,499.00
ORACLE	Proprietary	$ 4,350.00
HORTON-Works data platform	Open Source	Free

F. A comparative analysis based on scalability

Scalability is one of the criteria to measure the capability of a given software or system. Scalability is all about the expansion support of a given system when there is a need for expansion in demand.

If the scalability of these tools is analyzed; most of the RDBMSs are not horizontally scalable however they are vertically scalable till there maximum limitations. As a matter of facts; scalability of Big Data tools is not limited horizontally as well as vertically.

G. Comparative analysis based on Data visualization support

Visualizations help people see things that were not obvious to them before. Even when data volumes are very large, patterns can be spotted quickly and easily [14]**.** Its fact that the data visualization will make data analysis results to be presented in a best possible way as crystal clear. Now every data analytics tools and technologies are including this feature in there system.

Using RDBMS one can do analytics on them but can't visualize the result of the analysis, however using Big Data tools and technologies every data analysis can be supported by the data visualizations.

H. Comparative analysis based on ease of deployment

Deployment is the first and sometimes the only experience system administrators have with an application. Ease of deployment is a key consideration for any systems. Most of the time installation of software on windows operating system is easy and simple it's all about opening the executable file then following the prompt. However when it comes to installing software in non-windows operating system use needs a key knowledge working with the terminal and the command line scripting. Its fact that windows operating system have 55.42 % of users across the world and that of non-windows operating systems has 44.58 % of users [15].

Based on this fact and observation installing software like RDBMSs on windows machines can be an easy task even if it's difficult for normal users to install them on non-windows operating system. However in Systems like Big Data tools, it has been observed that it's very difficult to configure and deploy them on windows machines and also it's a tough task to install them on non-windows operating systems without having a full knowledge terminal.

I. Comparative analysis based on System Security

In the digital age keeping information secured is very challenging task due to multiple threats imposed on the information systems keeping or storing the data for instance in 2013 Kaspersky lab have announced 5,188,740,554 cyber-attacks, identifies 104,427 newly modified malicious programs, 1,700,870,654 attacks on online resources from online attackers and 3 billion malware attacks were found [16].

In keeping the privacy and security of the information RDBMSs and Big Data analytical tools have their own solution or counter measures to every treats posed on them, however having and implementing all this tight security measures even didn't kept the data from hackers. In this research study the security features of each tool from both RDBMS and Big Data analytical tools and technologies were analyzed based on the product owner specifications and it was founded that large volumes of RDBMSs have a high security measures deployed on them to keep the safety of the data. However if user see the security measures implemented in Big Data tools and technologies there is a high concern for security and privacy issue to be addressed in future.

Finally it was observed, analyzed and concluded that most of the RDBMSs have provided us a plenty of security features that can help to secure data and to protect it from authorized users, however most of the big data analytical tool challenges are confined to security and privacy issues only.

VI. DECISION SUPPORT METRICS FOR DATA DRIVEN ORGANIZATIONS

As stated in the objective, the research as a final contribution designed a suitable decision support metrics that can be used by Data analysts/ data scientists or decision makers in selecting data analytic technologies and tools. The decision support metrics is illustrated in table 6:

TABLE VI. SELECTION DECISION SUPPORT METRICS AND RECOMMENDATIONS

Selection parameters		RDBMS	Big Data Tools
Structure of data	Structured	Good	Good
	Semi-structured	Satisfactory	Good
	unstructured	Poor	Good
Resource consumption & execution time	Read operation	Good	Satisfactory
	Write operation	Satisfactory	Good
Scalability	Vertically	Good	Good
	Horizontally	Poor	Good
Cost		Satisfactory	Good
Ease of deployment		Good	Satisfactory
Data visualization		Satisfactory	Good
Large data set		Poor	Good
Security and privacy		Good	Satisfactory

VII. CONCLUSION

The main thrust of this research study started from the notion; if structured, unstructured and semi-structured data emerge with 3Vs then how to handle them efficiently. In case current technologies are not capable enough to handle; then what next? After the rigorous observation and analysis of different features of RDBMSs and Big Data, it is concluded that the major challenge in Big Data is storage capacity and it can be fulfilled by the Hadoop distributed file system (HDFS) and the analysis process can be handled by a Map and Reduce which can process data across different clusters in parallel manner. A metrics for purchase rent or deploy related decision support for data management or handling is also designed here. This 'Decision Support Metrics' can be used as an 'advisory base line' for selecting most fit tools from available in the market. This metrics presented seven parameters for the analysis. Based on these parameters; decision makers can select appropriate tools and technology for optimizing performance in desired domain of application. This base line metrics can be used by data scientists as a business intelligence support tool to select best fit tools from RDBMSs, DBMSs and Big Data technologies based on organizational needs.

REFERENCES

[1] "Edgar F. Codd," 20 05 2014. [Online]. Available: http://en.wikipedia.org/wiki/Edgar_F._Codd.

[2] "Big Data," Wikipedia, [Online]. Available: http://en.wikipedia.org/wiki/Big_data. [Accessed 23 05 2014].

[3] "Diffrenece between RDBMS and," Google, [Online]. Available: https://www.google.com.et/?gws_rd=ssl#q=difference+between+rdbms+and+big+data. [Accessed 15 07 2014].

[4] Intel, "Peer Research Big Data Analytics," Intel, 2012.

[5] N. F. H. M. K. Judith Hurwitz, Big Data for Dummies, New York: Willey, 2013.

[6] Global Pulse, "Big Data for Development: Challenges and Opportunities," Global Pulse, 2012.

[7] General Daynamics, "Current Data Security Issues of NoSQL Databases," General Daynamics, 2014.

[8] "Hadoop Wiki," Apache, [Online]. Available: http://wiki.apache.org/hadoop/Distributions%20and%20Commercial%20Support. [Accessed 11 07 2014].

[9] "Hadoop," Hortonworks, [Online]. Available: http://hortonworks.com/hadoop. [Accessed 10 07 2014].

[10] "Navicat Premum," Navicat Premum, [Online]. Available: http://www.navicat.com/products/navicat-premium. [Accessed 25 07 2014].

[11] "Vertualization," Tech Target, [Online]. Available: http://searchservervirtualization.techtarget.com/definition/virtualization. [Accessed 26 07 2014].

[12] "Using Windows Performance Monitoring Tool," TEchNet, [Online]. Available: https://technet.microsoft.com/en-us/library/cc749115.aspx. [Accessed 29 07 2014].

[13] "Big Data," SAS, [Online]. Available: http://www.sas.com/big-data/. [Accessed 10 06 2014].

[14] "Data Visualization what it is and why it is important," SAS, [Online]. Available: http://www.sas.com/en_us/insights/big-data/data-visualization.html. [Accessed 9 08 2014].

[15] "usage share of OS," Wikipedia, [Online]. Available: http://en.wikipedia.org/wiki/Usage_share_of_operating_systems. [Accessed 13 08 2014].

[16] V. Beal, "security (computer security)," Webopedia, [Online]. Available: http://www.webopedia.com/TERM/S/security.html. [Accessed 23 8 2014].

AUTHORS PROFILE

Alazar Baharu: Alazar Baharu is academic staff in AMiT CS&IT Department at AMU. He is having additional charge of Library Director in order to his benevolent dedication and expertise for digitalization of library operations and services. He has developed many significant automation systems and services which are successfully deployed at his working organization like Demographic Heath surveillance system for which he is actively supporting the system as Data Manager. He has sound knowledge of Big Data analytics for decision support services and Information System Science.

Durga Prasad Sharma: Confluence of the three, an Eminent Academician, Technology Expert and Rehabilitation Activist, Prof. Durga Prasad Sharma has an excellent record of achievements and contributions in India and abroad. His contributions to the field of technology based rehabilitation under UN Convention for PWDs, ILO & UNDP schemes have been incomparable in the world. He is associated with Computer Science Teachers Association (CSTA-ACM) for USA & Canada and International Fellow of several organizations like SIE-Singapore and IACSIT-China & USA. He is recipient of 46 National and International Awards and wide range of appreciations. For his notable contributions to the research and academia, twenty nine world scientific organizations /societies have offered him various types of Membership and Fellowships. He has written 21 Computer Science and IT books and published more than100 research papers/ article. He has guided 9 PhD scholars from USA, Fiji, Saudi Arabia and India.

Solving Word Tile Puzzle using Bee Colony Algorithm

Erum Naz [1], Khaled Al-Dabbas[2], Mahdi Abrishami[3], Lars Mehnen[4], Milan Cvetkovic [5]
University of Applied Science Technikum Wien
Vienna, Austria

***Abstract*—In this paper, an attempt has been made to solve the word tile puzzle with the help of Bee Colony Algorithm, in order to find maximum number of words by moving a tile up, down, right or left. Bee Colony Algorithm is a type of heuristic algorithms and is efficient and better than blind algorithms, in terms of running time and cost of search time. To examine the performance of the implemented algorithm, several experiments were performed with various combinations. The algorithm was evaluated with the help of statistical functions, such as average, maximum and minimum, for hundred and two-hundred iterations. Results show that an increasing number of agents can improve the average number of words found for both number of tested iterations. However, continuous increase in number of steps will not improve the results. Moreover, results of both iterations showed that the overall performance of the algorithm was not much improved by increasing the number of iterations.**

Keywords*—slide tile puzzle; artificial bee colony algorithm; swarm intelligence; artificial intelligence; fitness function; loyalty function; word tile puzzle; Bee colony optimization

I. INTRODUCTION

The prime inspiration to design any optimization algorithm is to simulate natural processes. Lots of algorithms have proved their inspiration from natural process such as simulated annealing (SA), genetic algorithms (GA) [1], ant colony optimization (ACO) [2], particle swarm optimization (PSO) and other Swarm Intelligences (SI) [3]. Swarm Intelligence is based on the collective behaviour of individuals in various decentralized systems. These decentralized systems are composed of physical individuals that communicate, cooperate, collaborate, and exchange information and knowledge among themselves to perform some tasks in their environment [4]. A detailed survey to related work is discussed in Section II.

The purpose of this paper is to develop a slide tile game. An attempt has been made to solve the word tile puzzle problem for the most optimum solution by taking an inspiration from natural processes. In the beginning, the board of the tile game is filled with random characters. The tile game is of size n x n, starting from 4 x4 (Figure 1), but in the end larger sizes should be solvable. By moving the tiles (up, down, right, left) a specific situation should be found, where the graph contains a maximum number of words (length 2 to n).

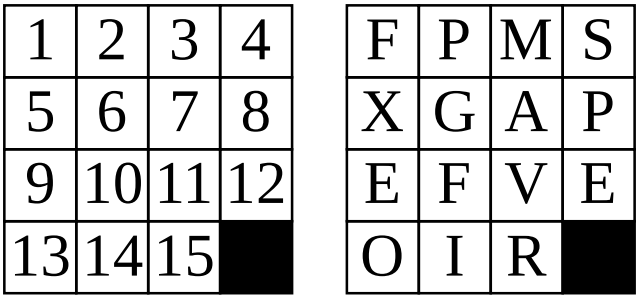


Fig. 1. 15-puzzle slide tile game and 15-word tile puzzle

This paper is organized as follows: Section I gives the brief introduction to the topic, Section II is about related work in the field of string matching, crossword and tile puzzles, Section III elaborates the implemented steps of bee colony algorithm, Section IV describes the statistical results, Section V is discussing the conclusion and finally section VI is future work.

II. LITERATURE REVIEW

Hua [5] implemented blind search and heuristic search algorithms to solve Eight-puzzle problem. In blind search Breadth-first and Depth-first and in heuristic search A* algorithm were implemented to find the optimal solution. The result showed that A* algorithm is more convenient, efficient and better in terms of running speed and cost of search time than blind search algorithms. Genetic algorithm and depth first algorithm was implemented to solve Japanese puzzles. Evaluation and comparisons of performance concluded that depth-first algorithm is faster for small size puzzles and genetic algorithm is better in large size puzzles. However, both methods are slow [6].

One of the recent optimizations algorithms replicating the intelligent natural behavior of honeybee's swarm is artificial bee colony (ABC) [10]. The popularity of ABC increased significantly in last years, the algorithm has been implemented in different field for example in [11] used ABC to solve complex network, while [12] combined Fuzzy logic with ABC in the optimization of parameters for an autonomous mobile robot control.

Author in [7] used the natural behaviours of the real honey bee to develop the artificial bee colony algorithm and resolve numerical optimization problems. In the Artificial Bee Colony algorithm (ABC), there are three types of bees, scout bees, employed bees and onlooker bees. The half of the total bees are initially scout bees. For each scout bee, a new food source position is produced. After producing new food source position, these scout bees become employed bees. Then these employed bees try to improve food sources by interacting with each other. The onlooker bees wait for the food sources positions by the employed bees in the hive. Employed bees share position information about the food sources, each onlooker bee picks up one of the food source positions and tries to improve the food source position.

Like ABC, the Bee Colony Optimization (BCO) algorithm, follows the way how honey bees in real-world nature look for food source, which makes this algorithm more effective, compared to lots of other stochastic random-search algorithms. So far, BCO is implemented in various real-life optimization problems, such as vehicle routing problem [8], the routing and wavelength assignment in all-optical networks, the traffic sensor location problems on highways, the static scheduling of independent tasks on homogeneous multiprocessor systems and disruption management in public transit [9].

III. METHODOLOGY

This section elaborates the step by step approach to solve word tile puzzle by using bee colony algorithm [8].

A. Step 1: Initial Input

The Bee Colony Algorithm starts with the initial inputs. These inputs are:

1) Agents: an entity that performs the activity – Artificial bees.

2) Number of Steps per Agents: number of shifts movements by each agents or distance covered during one trip.

3) Number of Iterations: number of trips made by each agents.

The number of agents, represent the number of employed artificial bees, looking for food source. In this case, the agents will look for the best state value of the board, which is the maximum number of words found in rows and columns. The number of steps implies how far the agents can go during one iteration (In Bee Colony Algorithm, the number of steps indicate, how far the bees can travel). The final input value is the number of iterations, which demonstrates the number of trips per each run of the program (The number of trips that each employed bee will perform in an assigned working period). It has to be taken into account that the output value of implemented program will highly depend on our initial inputs.

B. Step 2: Defining Initial State

In this step, the dimensions of the board are defined and initial position of the blank tile (starting point for agents/bees) is marked (tile in the last row and the last column in this case); random letters (alphabets) are assigned to the initial state of the board. The value of the initial state is calculated, by using the fitness function. Fitness Function counts the number of all existing words in the board by matching the input string to the imported dictionary. Word could be two letter long or maximum to the dimension of the board. Fitness function is calculated after each movement of the agent in the board. It is allowed, that more than one word is found in a row or a column and there can be a word inside another word. Word matching is from left to right and top to bottom (diagonal matching is not allowed).

In order to reduce the computational cost and improve performance, two constraints are applied to the dictionary. The first constraint is to eliminate the words longer than the dimension of the board and the second constraint omits the words which were not the part of randomly generated board in the initial state.

C. Step 3: Agents movement and state update

In this step, the movement of agents is assigned randomly and new state per each agent is generated.

Agents can move up, down, right and left inside the board. The last movement of the agents and their directions are stored to avoid a backward movement. The idea behind not letting agents to move backwards is to avoid unnecessary movements which will be not only time consuming, but they will also take some memory computation. Agent cannot move diagonally and beyond the walls of the board.

After defining initial input, states and keeping constraints in consideration, agents are ready to move and start searching for higher value state. After each step of each agent, the value state is calculated (using fitness function). If an agent reaches higher value of Fitness Function (number of words) than value of existing best state/states will be stored by updating archive of best states and previous best state/states will be deleted. If an agent reach equivalent value of Fitness Function as value of existing best solution or solutions its state will be added to other states with the same value in the archive.

D. Step 5: Agents' Loyalty Function

In this step, agent's loyalty is determined by using probability. The idea of considering loyalty comes from the natural processes inspiration. Studying the behaviour of bees for finding nectar, it has been observed that the successful bees in finding food sources will go back to the hive, performing the "Waggle dance", in order to convince other bees to get the same route as them and go for the same food source.

This behaviour in this algorithm, has been defined by loyalty function. After each step of each agent per each iteration, the state value of that specific agent will be calculated and compared to the state values of other agents. Certainly, there will be agents with higher state values, and the loyalty function will make other agents to decide, if they want to keep on their own route or they would rather change their routes, which show that there might be higher possibility of finding better state values.

1) Loyalty of Agent: To compute the loyalty of agents, following probability functions have been used [9].

$$P_n = e^{-\{(O_{max} - O_n)/k\}} \quad (1)$$

P_n: Probability of agent n to stay loyal

O_{max}: MAX (O_n) =1 , O_{max} is maximal normalized value of fitness functions of all agents
O_n: $(C_n – C_{min})/(C_{max} - C_{min})$
C_n is normalized value of fitness function for agent n
C_{min} is minimum value of all fitness functions
C_{max} is maximum value of all fitness functions
k: number of current step in current iteration

The methodology is based on the roulette game; a number, which is between 0 and 1, should randomly be chosen and if the value of this random number is less than P_n, then agent n is considered as loyal, otherwise, it is counted as disloyal and change its route for the agents with better state values.

Assigning new state to disloyal agents: The next step, will be for disloyal agents to choose a state of all possible loyal agents' states. However, there is the possibility that there might be more than one loyal agent, whose state, is a nominee of being chosen as the higher value state. Or there can be more than one disloyal agent, which can make a choice out of different states of nominate loyal agents. In this algorithm, probability functions in (2) has been implemented that compute the probability of being chosen as the higher state value agent by disloyal agents:

$$P_A = \frac{O_r}{\sum_{r=1}^{R} O_r} \quad (2)$$

P_A Probability of loyal agents to be chosen by disloyal agent
r: Loyal agent
R: Total number of Loyal Agents
O_r : Normalized Fitness Function value of loyal agent
$\sum_{r=1}^{R} O_r$: Sum of normalize Fitness Function values of all loyal agents

In this formula, the normalized fitness function value of each loyal agent is divided by the sum of fitness function values of all loyal agents. Finally, to each disloyal agent, a state of a loyal agent would be assigned, based on the outcomes of the probability functions.

E. Step 6: Best Value State (Optimal Solution)

After the last iteration is done, the best value state is printed from archive of the best state or states that is/are archived during the whole search process, and the existing words in the board will be presented.

IV. Results

In order to shape our results, number of experiments were performed. Implemented solution is tested for number of steps, number of agents and number of iteration. To determine the effect of each variable, they were tested separately. To achieve this, variables other than test variables were kept constant and results were calculated, using fitness function, by changing the value of test variables.

A. Evaluation measurements

Basic functions of statistics such as average, maximum and minimum are used to calculate the results.

1) Average: average number of words found per 30 testing experiments for defined number of agents (2, 4, 8, 12, 16, 20, 30) and number of steps (10, 20, 30, 40) with fixed number of iteration.

2) Maximum: the highest value of the state of the board per 30 testing experiments with fixed number of iterations.

3) Minimum: the lowest value of the state of the board per 30 testing experiments with fixed number of iterations.

The main dictionary consisting of 22,353 words was used for testing, however it is reduced to 1,359 words after passing through the dictionary reduction function. 2, 640 experiments were performed including 100 and 200 iterations, 2, 4, 8, 12, 16, 20 and 30 agents and 10, 20, 30 and 40 steps. Detailed results are presented below.

B. Results for 100 iterations

1) Varying number of steps for 100 iterations

To examine the effect of the number of steps, 660 experiments were executed, in which the number of steps were changed while keeping the number of agents and the number of iterations constant. The number of steps started from 10 extending to 20, 30 and 40.

Table I shows the test results for 100 iterations, varying number of steps, starting from 10 and then 20, 30 and 40 for 2, 4, 8, 12 and 16 agents.

TABLE I. Varying Number of Steps for 100 Iterations

Varying number of steps for 100 iterations		**Number of steps**			
		10	*20*	*30*	*40*
2 Agents	Average number of words	27.57	28.5	28.57	28.93
	Maximum number of words	33.00	32.00	32.00	32.00
	Minimum number of words	23.00	24.00	25.00	25.00
4 Agents	Average number of words	27.7	29.93	30.30	30.23
	Maximum number of words	31.00	34.00	36.00	36.00
	Minimum number of words	24.00	26.00	26.00	25.00
8 Agents	Average number of words	29.13	30.97	30.50	30.47
	Maximum number of words	33.00	34.00	35.00	35.00
	Minimum number of words	25.00	28.00	27.00	27.00
12 Agents	Average number of words	28.53	31.17	31.43	31.90
	Maximum number of words	35.00	35.00	38.00	36.00
	Minimum number of words	23.00	27.00	27.00	27.00
16 Agents	Average number of words	29.27	32.13	32.57	32.10
	Maximum number of words	34.00	36.00	36.00	40.00
	Minimum number of words	26.00	28.00	29.00	27.00

There is an observable increase in average, maximum and minimum number of words when increasing number of steps from 10 to 20. However, after first 20 steps, there is no noticeable improvement and minimum number of words found, did not progress.

Fig. 2, shows the graphical overview of results obtained by varying number of steps for 100 iterations.

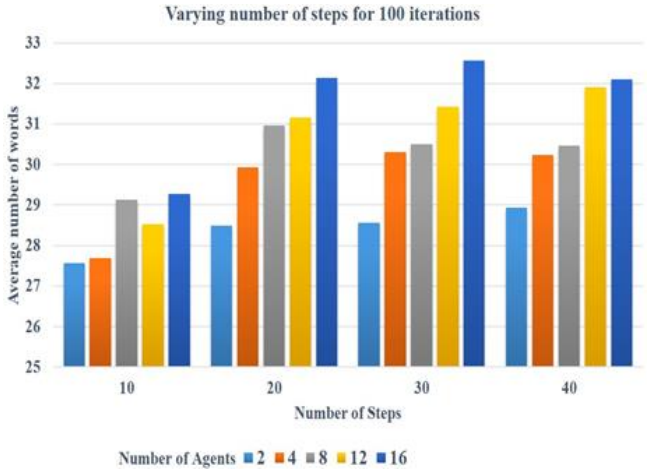


Fig. 2. Varying number of steps for 100 iterations

2) Varying number of Agents for 100 iterations

To examine the effect of the number of agents, 660 experiments were executed, in which the number of agents were changed while keeping the number of steps and the number of iterations constant. The number of agents started from 2, extending to 4, 8, 12, and 16.

Table II, shows the test results for 100 iterations, varying number of agents, starting from 2, 4, 8, 12 and 16 for 10, 20, 30 and 40 steps.

It was noticed that increasing the number of agents improves the average number of words. However, the maximum and the minimum number of words were not much effected.

TABLE II. VARYING NUMBER OF AGENT FOR 100 ITERATION

Varying number of Agent for 100 iteration		**Number of Agents**			
		2	*4*	*8*	*12*
10 Steps	Average number of words	27.57	28.5	28.57	28.93
	Maximum number of words	33.00	32.00	32.00	32.00
	Minimum number of words	23.00	24.00	25.00	25.00
20 Steps	Average number of words	27.7	29.93	30.30	30.23
	Maximum number of words	31.00	34.00	36.00	36.00
	Minimum number of words	24.00	26.00	26.00	25.00
30 Steps	Average number of words	29.13	30.97	30.50	30.47
	Maximum number of words	33.00	34.00	35.00	35.00
	Minimum number of words	25.00	28.00	27.00	27.00
40 Steps	Average number of words	28.53	31.17	31.43	31.90
	Maximum number of words	35.00	35.00	38.00	36.00
	Minimum number of words	23.00	27.00	27.00	27.00

Fig. 3, shows the graphical overview of results obtained by varying number of Agents for 100 iterations.

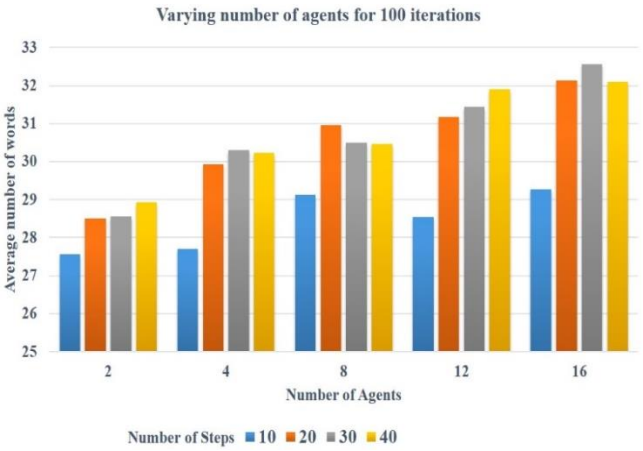


Fig. 3. Varying number of agents for 100 iterations

C. Results for 200 iterations

1) Varying number of steps for 200 iterations

Table III shows the varying number of steps for 200 iterations by keeping agents constant. The number of steps started from 10, extending to 20, 30 and 40.

TABLE III. VARYING NUMBER OF STEPS FOR 200 ITERATIONS

Varying number of Steps for 200 iterations		**Number of steps**			
		10	*20*	*30*	*40*
2 Agents	Average number of words	28.30	29.50	29.90	30.40
	Maximum number of words	34.00	35.00	35.00	34.00
	Minimum number of words	23.00	26.00	26.00	27.00
4 Agents	Average number of words	28.50	31.10	30.80	31.60
	Maximum number of words	32.00	35.00	35.00	36.00
	Minimum number of words	25.00	27.00	27.00	28.00
8 Agents	Average number of words	30.20	31.40	32.00	32.20
	Maximum number of words	34.00	35.00	36.00	36.00
	Minimum number of words	26.00	26.00	29.00	29.00
12 Agents	Average number of words	30.40	31.90	32.30	32.60
	Maximum number of words	36.00	36.00	37.00	37.00
	Minimum number of words	25.00	29.00	27.00	29.00
16 Agents	Average number of words	29.80	32.40	33.50	33.20
	Maximum number of words	36.00	36.00	37.00	37.00
	Minimum number of words	25.00	27.00	27.00	30.00

Like 100 iterations, with 200 iterations, an observable increase in average number of words was found, when increasing number of steps from 10 to 20. Afterwards, there is no noticeable improvement for average number of words. However, for maximum and minimum number of words, there is a variation in results as compared to 100 iterations. Minimum number of words improved, when moving from 30 to 40 steps which was not the case in 100 iterations.

Fig. 4, shows the graphical overview of results obtained by varying number of steps for 200 iterations.

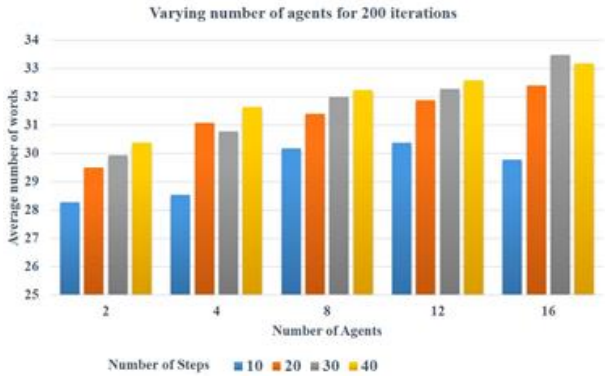


Fig. 4. Varying number of steps for 200 iterations

2) Varying number of Agents for 200 iterations

Table IV is presenting the results for varying number of agents for 200 iterations by keeping steps constant. The number of agents started from 2, extending to 4, 8, 12, and 16.

TABLE IV. VARYING NUMBER OF AGENT FOR 200 ITERATION

Varying number of Agent for 200 iteration		Number of Agents			
		2	*4*	*8*	*12*
10 Steps	Average number of words	28.30	28.50	30.20	30.40
	Maximum number of words	34.00	32.00	34.00	36.00
	Minimum number of words	23.00	25.00	26.00	25.00
20 Steps	Average number of words	29.50	31.10	31.40	31.90
	Maximum number of words	35.00	35.00	35.00	36.00
	Minimum number of words	26.00	27.00	26.00	29.00
30 Steps	Average number of words	29.90	30.80	32.00	32.30
	Maximum number of words	35.00	35.00	36.00	37.00
	Minimum number of words	26.00	27.00	29.00	27.00
40 Steps	Average number of words	30.40	31.60	32.20	32.60
	Maximum number of words	34.00	36.00	36.00	37.00
	Minimum number of words	27.00	28.00	29.00	29.00

Here again, like 100 iterations, with 200 iterations, increasing the number of agents, improve the average number of words and increase in maximum number of words. However, the behaviour of minimum number of words is mixed.

Fig. 5, shows the graphical overview of results obtained by varying number of Agents for 200 iterations.

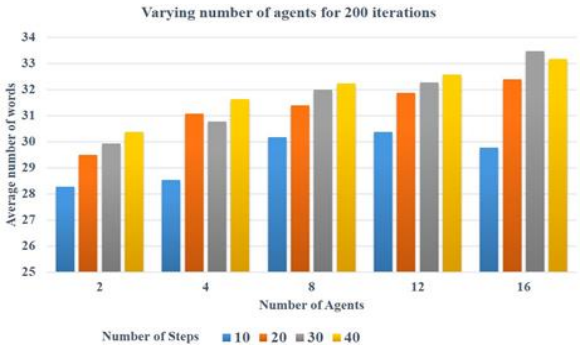


Fig. 5. Varying number of agents for 200 iterations

V. DISCUSSION/ANALYSIS

The performance of Artificial Bee Colony Algorithm is measured by executing larger number of experiments with multiple combinations. A huge variation in results was witnessed due to random behaviour of agents (bees). Average, maximum and minimum number of words found increased, while increasing number of steps from 10 to 20 for both 100 and 200 iterations. However, there is no significant change noticed by increasing steps from 20 to 30 or onwards. Experiments with 20 and 30 agents were also performed, due to high insignificance, those results are not presented in this paper. For 100 iterations, best result is found for 16 agents and 30 steps (average number of words=32.57), following with 20 steps and 16 agents (average number of words=32.13), whereas for 200 iterations best result found was for 16 agents and 30 steps (average number of words=33.50) followed by 16 agents and 40 steps (average number of words=33.20).

By comparing results for 100 and 200 iterations, it is observed that overall performance of the algorithm is not much improved by increasing number of iterations. However, the behaviour of algorithm for 100 and 200 iterations is almost the same for average number of words, whereas for maximum and minimum number of words, it is diverse.

Fig. 6, shows the overall improvements of best result for 100 iterations. 100% of improvements are achieved in 96 iterations. The graph shows that for first 16 iterations, there is almost 70% improvement in results and the remaining 30% improvement covers 80 iterations.

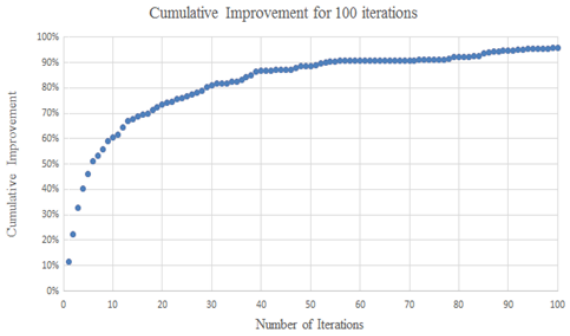


Fig. 6. Cumulative improvement for best result of 100 iterations

Fig.7, shows the overall improvements of best result for 200 iterations. 100% of improvements are achieved in 194 iterations. The graph shows that for first 16 iterations there is almost 61% improvement in results and the remaining 39% improvement covers 178 iterations.

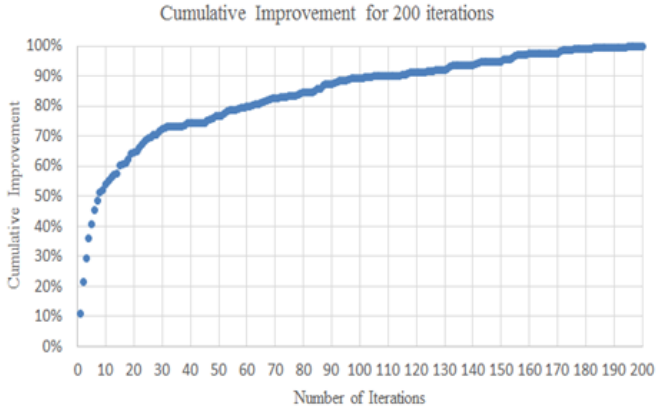


Fig. 7. Cumulative improvement for best result of 200 iterations

Furthermore, in this experiment the time needed to achieve the solution was not taken into consideration as nowadays modern computer have different computing power, and the goal was to achieve the best result regardless of time. Although, the computation time for best results (only) were calculated and it is analysed that time to compute results for 200 iterations is much higher as compared to 100 iterations. Moreover, the improvement in overall results for 200 iterations are not that high. Table V, shows the exact time for best results.

TABLE V. TIME CONSUMPTION

Time Consumption		**Number of steps**		
		20	*30*	*40*
16 Agents	100 Iteration	1:43	2:24	3:04
	200 Iterations	3:56	5:01	6:29

Fig. 8, shows the bar chart for time consumption for best results.

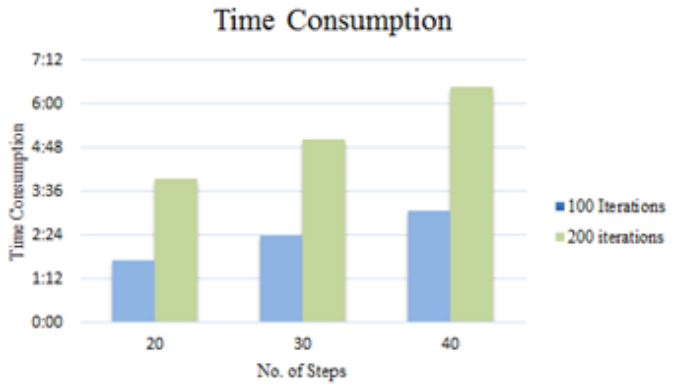


Fig. 8. Cumulative improvement for best result of 200 iterations

VI. CONCLUSION

A wide range of algorithms are inspired by natural processes proved to be successful in solving complicated optimization problems. Bee colony is considered as a class of swarm intelligence technique, where the corporation between different gents increase the efficiency and increase the probability of achieving better results, which cannot be achieved by individual agents. In this paper, word tile puzzle has been analysed using one of the heuristic algorithm named as Bee Colony Algorithm because of the way that artificial bees can communicate with each other and exchange information, making this method, not only fast, but also statistically optimal. Results showed that best solution could be achieved by increasing number of agents, nevertheless results are not improving with increasing number of iterations and steps continually. Furthermore, huge amount of time is required to run high number of iterations and results which has very nominal effect on results.

VII. FUTURE WORK

In future, other heuristic and blind algorithms can be implemented together for the word tile puzzle to compare the efficiency of each algorithm.

ACKNOWLEDGMENT

We would like to express our deep sense of gratitude to our mentor Dr. Lars Mehnen for guiding and encouraging us and one of our colleague Mr. Jakub Stok for his support and help.

REFERENCES

[1] J. Holland, Adaptation in Natural and Artificial Systems, University of Michigan Press, 1975.

[2] W. Gao, "Study on Immunized Ant Colony Optimization," in Natural Computation, 2007. ICNC 2007. Third International Conference, Haikou, 24-27 Aug. 2007.

[3] D. Teodorović, "BEE COLONY OPTIMIZATION: RECENT DEVELOPMENTS AND APPLICATIONS," "Mircea cel Batran" Naval Academy Scientific Bulletin, vol. XVIII, no. 2, pp. 225-235, 2015.

[4] D. Teodorović, "Bee Colony Optimization (BCO)," in Innovations in Swarm Intelligence, Springer Berlin Heidelberg, 2009, pp. pp 39-60.

[5] R. Shi, "Searching Algorithms Implementation and Comparison of Eight-puzzle Problem," in International Conference on Computer Science and Network Technology , Harbin, 24-26 Dec. 2011.

[6] W. Wiggers, "A Comparison of a Genetic Algorithm and a Depth First Search Algorithm Applied to Japanese Nonograms," in Twente Student Confer D. Karaboga, "D. Karaboga, An Idea Based on Honey Bee Swarm for Numerical Optimization," Technical Report – TR06, Turkey, 2005.ence on IT, Jun. 2004.

[7] Š. Milica and T. Dušan, Computational Intelligence in traffic, University of Belgrade - Faculty of Transport, 2012.

[8] M. Nikolić and D. Teodorović, "Empirical study of the Bee Colony Optimization (BCO) algorithm," Expert Systems with Applications, vol. 40, no. 11, p. 4609–4620, September 2013.

[9] M. Nikolić and D. Teodorović, "Empirical study of the Bee Colony Optimization (BCO) algorithm," Expert Systems with Applications, vol. 40, no. 11, p. 4609–4620, September 2013.

[10] B. B. D. Karaboga, "On the performance of artificial bee colony (ABC) algorithm," Applied Soft Computing, vol. 8, no. 1, pp. 687-697, 2008.

[11] D. D. Magdalena Metlicka, "Complex Network based Adaptive Artificial Bee Colony algorithm," in IEEE Congress on Evolutionary Computation (CEC), 2016.

[12] Leticia Amador-Angulo, "A Generalized Type-2 Fuzzy Logic System for the dynamic adaptation the parameters in a Bee Colony Optimization algorithm applied in an autonomous mobile robot control," in IEEE International Conference on Fuzzy Systems (FUZZ-IEEE), 2016.

A Novel Approach to Automatic Road-Accident Detection using Machine Vision Techniques

Vaishnavi Ravindran
Department of Computer Science
R.V. College of Engineering
Bangalore, India

Lavanya Viswanathan
Department of Computer Science
R.V. College of Engineering
Bangalore, India

Dr. Shanta Rangaswamy
Department of Computer Science
R.V. College of Engineering
Bangalore, India

***Abstract*—In this paper, a novel approach for automatic road accident detection is proposed. The approach is based on detecting damaged vehicles from footage received from surveillance cameras installed in roads and highways which would indicate the occurrence of a road accident. Detection of damaged cars falls under the category of object detection in the field of machine vision and has not been achieved so far. In this paper, a new supervised learning method comprising of three different stages which are combined into a single framework in a serial manner which successfully detects damaged cars from static images is proposed. The three stages use five support vector machines trained with Histogram of gradients (HOG) and Gray level co-occurrence matrix (GLCM) features. Since damaged car detection has not been attempted, two datasets of damaged cars - Damaged Cars Dataset-1 (DCD-1) and Damaged Cars Dataset-2 (DCD-2) – was compiled for public release. Experiments were conducted on DCD-1 and DCD-2 which differ based on the distance at which the image is captured and the quality of the images. The accuracy of the system is 81.83% for DCD-1 captured at approximately 2 meters with good quality and 64.37% for DCD-2 captured at approximately 20 meters with poor quality.**

Keywords*—Feature extraction; Image denoising; Machine vision; object detection; Supervised learning; Support vector machines

I. Introduction

A novel approach using image processing and machine learning tools to detect damaged cars from static images, which can be used to detect a road accident automatically is proposed. Detection or recognition of damaged cars falls under the category of object detection. Object detection or recognition using Machine Vision is achieved in two stages [1]. The first stage is feature extraction in which features common to instances from the object category are extracted from the corresponding images. The second stage includes training of a learning model like Support Vector Machines [2], Neural Networks [3] and AdaBoost [4] with the extracted features [1]. Principles used in most object detections do not work for detecting damaged instances of an object category since the damaged instances do not have the commonly extracted characteristics like shape, edges, Histogram of Gradients in common.

In this paper, a supervised learning method that detects damaged cars by making using of two facts – the state-of-the-art vehicle detection classifiers will not detect a damaged car and most damaged cars still have one or more car parts intact, is proposed.

The experimental results obtained show that the proposed approach gives promising results when tested on two different datasets of damaged cars which differ based on the quality, distance of the camera from the object and number of objects in an image. The two datasets were compiled for the sake of the project from various sources. The proposed method can be extended to other vehicles as a part of future work.

The work done includes three contributions. The first contribution includes proposing a novel approach to automatic road accident detection. The second contribution includes a supervised learning method that detects damaged cars from static images, a class of object that has not been detected so far using the techniques of machine vision. The third contribution includes the release of two public datasets of damaged cars- Damaged Cars Dataset-1 (DCD-1) and Damaged Cars Dataset-2 (DCD-2).

This paper is organized as follows. Section II describes related work. The proposed method is explained in Section III. The experimental results and analysis for two different datasets are presented in Section IV. Section V presents Conclusions.

II. Related Work

The Global status report on road safety 2015 [5] shows that the total number of deaths caused due to road accidents is at 1.25 million a year. One of the main reasons for fatalities from these accidents is a delay in reporting the accidents to near-by emergency health centres and delay in an ambulance reaching the accident location. Such a delay can be can be reduced if there is automatic detection and reporting of the accidents to emergency help centres. Most of the prevalent state-of the-art methods use sensor technology to detect road accidents. In [6],[7] the use of sensors present inside the vehicle including accelerometers, GSM and GPS modules is made to detect unusual movements and angles of the car to indicate an accident. In [8] the use of sensors like magneto resistive sensors is made outside the vehicle, installed on the roads. In [9], MMA621010EG is a proven special car accident sensor which is integrated XY-axis accelerometer and built-in serial peripheral interface SPI bus. The variations from this sensor is detected and through the GPS software fitted in the vehicle communication is made with the satellite. The latitude and longitude values are sent to the centralized server for contacting the emergency service. The drawback with using

sensor technology is that the sensors can get damaged in the accident. One way to overcome this drawback is by making use of the surveillance cameras installed in traffic junctions and highways. The static images or video footage from these cameras can be used in detecting the presence of a damaged vehicle which in turn indicates the occurrence of an accident. A new approach based on vision based object detection is presented in this paper which detects the presence of a damaged car from static images.

As seen in [10][11] detection of damaged buildings and roads involves working with images of the object prior and post damage. This however cannot be applied to damaged car detection for the purpose of accident detection.

III. PROPOSED METHOD

The proposed method is a supervised learning one which works as a binary classifier distinguishing between images containing a damaged car as class 1 and images not containing it as class 0. Instances of damaged cars do not have anything in common due to loss of shape, edges and intensity gradients. Hence using the usual vision-based object detection methods where HOG, Haar, Gabor and SURF features are used to train SVM, AdaBoost and neural networks [12] will not achieve detection. However, a feature that most damaged cars do have in common is the presence of at least one car part. Our classifier is based on this fact. But training a classifier that detects car parts alone will not achieve successful detection of damaged cars since cars without any damage also show the presence of car parts. Hence there is a need to differentiate between cars that are damaged and cars that are not in addition to the step involving detection of car parts. When the state-of-the-art classifiers developed so far [12] for detection of vehicles and cars in specific were tested, results showed that they failed to detect most damaged cars. This is the second fact that forms the basis of the method developed in this project.

The input images can be divided into three types - images of damaged cars (type 1), images of undamaged cars (type 2) and images of all other objects and scenes (type 3). The classifier built should now work as a binary classifier which distinguishes between images containing a damaged car (type 1) as class 1 and images not containing it (type 2 and type 3) as class 0.

The realization of the system is done in three different stages which are combined into a single framework in a serial manner as shown in Fig. 1. Prior to the first stage is the pre-processing of images. The first stage is the vision based detection of undamaged cars using a SVM trained with HOG [26]. The SVM works as a binary classifier detecting the presence of a car without any damage and thus separates type 2 (class 1) from type 1 and type 3 (class 0). Since damaged cars are very close in appearance to undamaged cars results showed stage 1 misclassifies a considerable number of damaged cars as class 1, that is, as undamaged cars. Hence in order to improve the performance of stage 1, a binary classifier that detects the presence of damaged texture from all images classified as undamaged cars in the previous stage was introduced, this is stage 2. It reduces the number of false positives from stage 1. The detection of damaged texture at this stage is done by training a SVM with GLCM features. The third stage is now used to separate type 1 from type 3 by using a car parts detector which consists of three separate binary classifiers, each detecting the presence of one car part. Each of the binary classifiers at this stage is a SVM trained with HOG features of the corresponding car part. The three car parts considered are wheel, headlight and hood. The method does not work in the cases where the car is damaged to an extant where it has none of the car parts considered, this serves as the limitation of the proposed method. It can be overcome in the future by increasing the number of car parts detected.

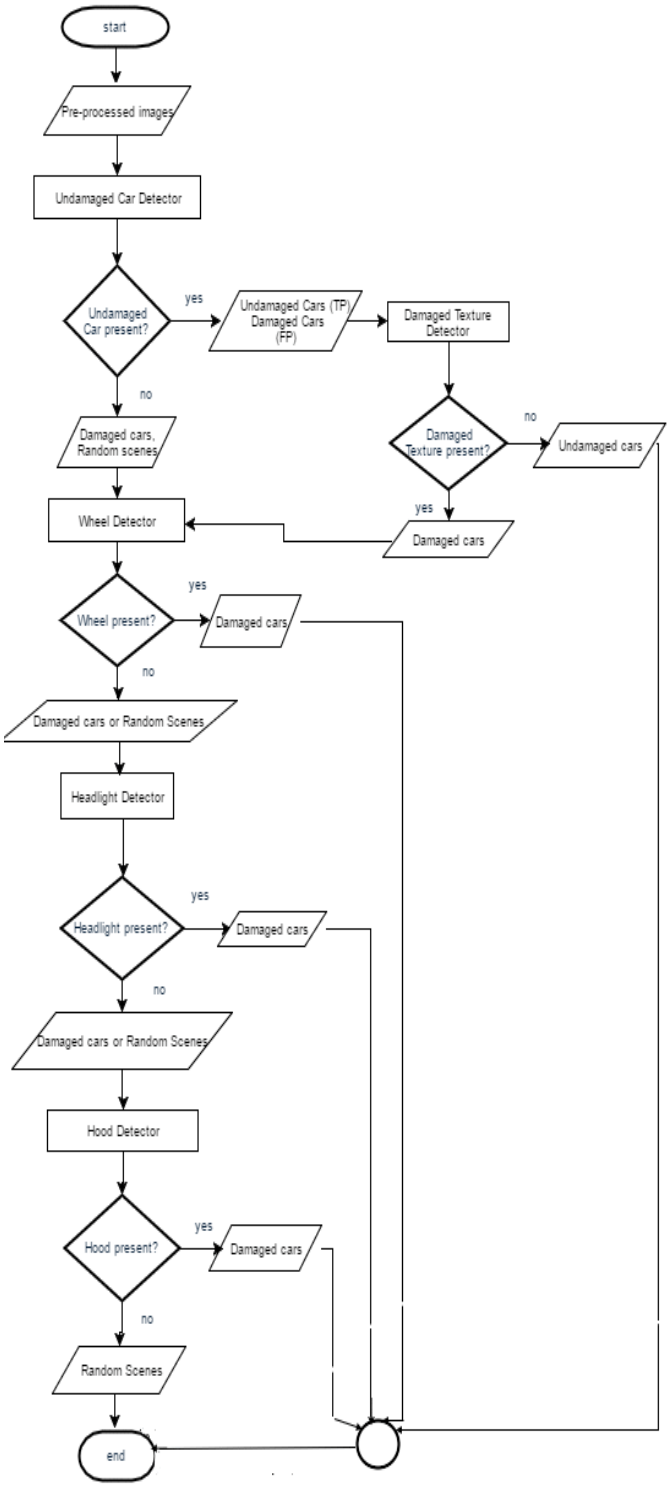


Fig. 1. Flow chart of the proposed method

A. Pre-processing

This is the first step that is carried in order to enhance the image and remove distortions like noise from the image. De-noising was achieved by using a Median filter [13]. The images are then resized to 256×256 for the next stage and converted to .JPEG format.

B. Undamaged Car Detector

Histogram of (HOG) features extracted from the training dataset is used to train a Support Vector Machine (SVM) which is a binary classifier based on supervised learning is employed in the first stage.

HOG is a feature descriptor introduced by Dalal et al.[14]. It is global feature and not a collection of many local features, that is, the object is described by one feature vector and not many feature vectors representing different parts of the object. The training images are resized to 256×256 and converted to gray scale. The HOG descriptor is then extracted from these 256×256 images, where 4×4, 8×8 or 16×16 pixels per cell are considered known as the cell size and from each cell the gradient vector at each pixel is calculated and put into a histogram with 8 bins. The histogram has 20 degrees in each bin and ranges from 0 to 180 degrees in total. Each gradient vector's magnitude is put into the histogram of bins and each value is split between the two closest bins. The reason behind the histogram is quantization and to reduce the number of values. Apart from this, it also generalizes the values in a cell.

The next step is normalizing the histograms to make them invariant to changes in illumination. When any vector is divided by its magnitude, it is said to be normalized. A gradient vector of a pixel in invariant to addition, subtractions and multiplications and hence the histograms can be normalized. This is taken one step further and instead of normalizing the histograms individually, the cells are grouped into blocks of 1× 1, 2×2 or 4×4 cells known as the block size and all the histograms in the block are normalized together. The reason behind this type of block normalization is that since changes in contrast occur in smaller regions of the image than compared to the entire image, normalizing in smaller blocks is more effective.

After extracting the HOG feature vectors, they are used to train an SVM model [15]. SVM is a binary linear classifier used for supervised learning. The SVM divides the training data into two classes by constructing a maxim margin hyperplane such that this plane or surface has the maximum distance from the closest points in each training set called support vectors.

In order to get the optimal parameters for the SVM used to classify undamaged cars, a Grid Search [16] is performed which takes all the combinations of the parameters and finds the optimal set using cross-validation [16]. The parameters that are involved in Grid Search include the type of kernel (linear or RBF), value of C and Gamma.

C. Damaged Texture Detector

This is the second stage which improves the performance of the first stage by reducing the number of false positives from the first stage since the undamaged car detector detects a few damaged cars as undamaged cars due to their similarity. This module is implemented by training a SVM with texture features. Two types of texture features, Local Binary Patterns (LBP) and Grey Level Co-occurrence Matrix (GLCM) were extracted and it was seen that GLCM had a better performance.

Local Binary Patterns (LBP) [17][18] is a texture descriptor in which the image is divided into cells with each cell containing a fixed number of pixels. Each pixel in a cell has its value compared to its 8 neighboring pixels (north, south, east, west, north-west, north-east, south-west and south-east). If the neighbor's value is greater, then a 0 is assigned else a 1. Each pixel thus generates an 8-digit value and all such values are used to compute a histogram. Each cell's histogram is then normalized and concatenated to form the feature vector.

Gray-level co-occurrence matrix (GLCM) is method of extracting texture features based on the spatial relationship between pixels. In 1973, the method was proposed by Haralick et al., [19]. In this method pairs of pixels are considered in specific spatial position and values and a matrix is constructed, from this matrix various statistical texture characteristics are extracted. The texture characteristics that were considered for this work were dissimilarity, energy, contrast, homogeneity, correlation and Angular Second Moment (ASM). These texture characteristics were combined together to form a feature vector.

D. Car Parts Detector

This is the final stage and consists of three SVM classifiers each trained to detect one car part. The three car parts considered are wheel, headlight and hood. The steps to train and test the SVM model with Histogram of Gradient Descent features are similar to the methods employed in undamaged car detector. The three classifiers are cascaded with each other to segregate images of type 1 (damaged cars) and type 3 (random scenes) where the detection of one of the parts leads to classification of the image as a damaged car (type 1). An image is passed to the next classifier in this stage only if it is classified as negative by the previous classifier. An image that passes through and classified as negative by all three classifiers are classified as type 3 images.

E. Working of the system

A sliding window scans the image at each of the three stages and it is these scanned parts from one stage that are sent to the next stage. If at least one scanned part from an image results in the presence of a damaged car after passing through all three stages then the image to which it belongs is classified as containing a "damaged car". However, if all the scanned parts of an image are classified as a negative case of a damaged car, then the image is classified as negative case of containing a "damaged car".

The input image is first pre-processed by being converted to grayscale and resized to 256×256. It is then passed through the undamaged car detector and a sliding window of size 128×128 slides across the image at various scales. The 128×128 scanned parts that are classified as positive by the undamaged car detector are sent to the damaged texture detector and the ones classified as negative to the car parts detector. The damaged texture detector uses a sliding window of size 50×50 across the input images it receives. If a 50×50 scanned part shows the presence of damaged texture the corresponding 128×128 image is sent to the car parts detector. Similarly each classifier in the car parts detector uses a sliding window of size 50×50 and detection of a car part results in classifying the corresponding 128×128 input image as damaged. Finally, the original sized image to which this 128×128 scanned part belongs, is said to contain a damaged car.

If a damaged car is detected after these three stages an alert is sent to the nearest emergency help centre.

IV. Experimental Results and Analysis

In this section, the experimental datasets used, the setting of various parameters, experimental results and analysis are discussed.

A. Experimental Datasets

The dataset used for each of the three stages is discussed in this section. Two different data sets have been used for testing of the overall system.

The undamaged car detector uses 250 images of undamaged cars from [20] as the positive training data set, an example is shown in Fig. 2(a). From this data set only the portion of the images containing cars has been extracted using an object marker utility to extract the region of interest. 250 images of random scenes, that is, an image without the presence of a car, was taken from [21][22] as the negative training data set, an example is shown in Fig. 2(b). It was noticed that the performance at this step in regard to classifying damaged cars as negative cars was not satisfactory and hence 125 images of random scenes along with 125 images containing damaged cars only was added to the negative training data set from [23], an example is shown in Fig. 2(c) and available at [27] This improved the detection accuracy and the ability of the system to classify damaged cars as negative cars. Images of damaged cars were taken from the CIREN database [23]. The positive and the negative training data was resized to 256×256 and converted to grey scale.

Damaged texture detector uses 250 images of damaged cars from [23] made available at [28], as positive training data and 250 images consisting of undamaged cars from [20] and random scenes from [21][22] (that have not been used in the previous stage) as negative training data. The positive images are cropped to contain only the damaged parts using an object marker utility.

The car parts detector has three different classifiers for three car parts -wheel, headlight, hood- and each classifier uses 250 images from [20] of that part as training data as shown in Fig. 2(d), 2(e) and 2(f) respectively. 250 images of undamaged cars from [20] and random scenes from [21][22] are used as negative training data (that have not been used in the previous stages).

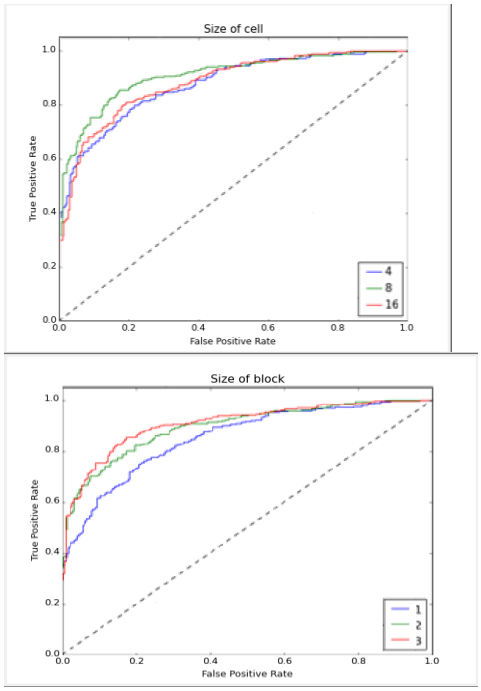
Fig. 2. Example images from the training and test datasets used. (a) and (h) are examples from the dataset used for undamaged cars, (b) from the dataset used for random empty scenes, (c) and (g) from DCD-1 and DCD-2, (d),(e) and (f) from datasets of the wheel, headlight and hood respectively

The test data for the overall system and each classifier was compiled and is called DCD-1 and DCD-2 as follows.

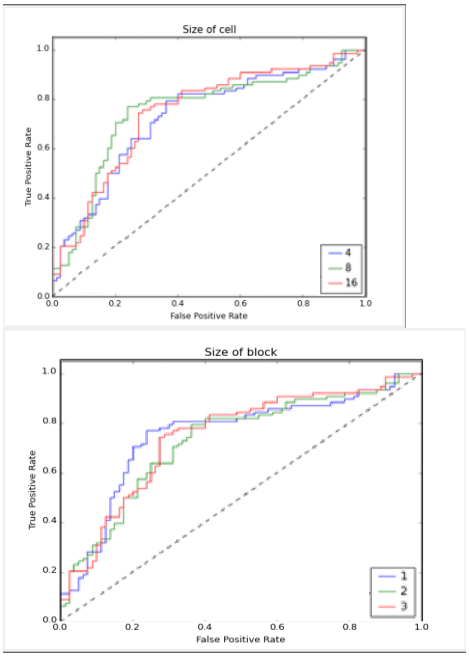
DCD-1

The first set- DCD-1- contains 300 images of damaged cars taken from [23] as the positive test data and 150 images of undamaged cars and 150 images of random scenes from [20][21][22] as negative test data (images that have not been used to train the model). The positive dataset contains images of individual damaged cars captured from different views (at approximately 2m from the damaged car) as shown in Fig. 2(c), available at [29] and the negative dataset contains images of individual undamaged cars captured from different views (at approximately 2m from the undamaged car) and images of empty scenes as shown in Fig. 2(a) and 2(b) respectively.

DCD-2- Since a part of the logic the system used is based on the difference between damaged and undamaged cars, images containing both types of cars were also considered, this forms the second set- DCD-2 - which contains images taken in real-time from surveillance cameras positioned on the sides of a road/highway with 80 images from [24] containing multiple damaged cars along with undamaged cars in a traffic scene as the positive test data set (at approximately 20m from the scene) as shown in Fig. 2(g) and available at [30]. 80 images from [24] containing multiple undamaged cars are used as the negative dataset (at approximately 20m from the scene) as shown in Fig. 2(h) and available at [31]. The images from DCD-2 were used for testing in an attempt to validate the working of the system in a realistic scenario since these images are of poorer quality, have higher changes in illumination and contrast and have more than one object present in the images.



(a)



(b)

Fig. 3. Results for various HOG parameter settings. (a) and (b) are the cell and block setting for DCD-1 and (3) and (4) are cell and block setting for DCD-2

B. Parameter setting

For the undamaged car detector, the optimal parameter for HOG feature extraction are evaluated for both datasets separately. The parameters are tested and optimized for HOG descriptor using ROC curves [25]. The parameters tested are size of the cell and size of the block where size of cell is the number of pixels contained in a cell and size of block is the number of cells contained in a block. The range of values considered for size of cell were 4×4, 8×8 or 16×16 pixels and the range considered for size of block were 1×1, 2×2, 3×3. The default parameters for HOG which are set while changing only one of the parameters are as follows:

- Cell size = 8,
- Block size=3
- number of bins = 9
- Step size= 10×10
- Minimum window size= 128

Fig. 3 shows that for DCD-1 when the different parameters for HOG were tested cell size 8 ×8 gave the best performance and for block size 3×3 gave the best performance and for DCD-2 cell size 8 ×8 and 16×16 and block size 1×1 and 3×3 gave the best performance. 8×8 was chosen as cell size and 3×3 as block size.

TABLE I. SVM PARAMETER RESULTS FOR DCD-1

Classifier	Kernel	C	Gamma
Undamaged Car	RBF	10	0.01
Damaged Texture	RBF	1	0.001
Wheel	RBF	10	10
Headlight	RBF	10	1
Hood	RBF	1	1

SVM: Grid search along with K-fold cross validation [16] is used for choosing the optimal parameters for SVM for each of the five classifiers used and for both Datasets. A Classification report is generated for the best parameter setting. The parameters that are tested for SVM are the type of kernel, C and Gamma values. Linear and RBF were the two type of kernels considered and for the value of C a range from 10^k for k € { -7,…..,7} and Gamma ranges from 10^k for k € { -6,…..,1}. The default parameters for SVM which are set while changing only one of the parameters are

TABLE II. SVM PARAMETER RESULTS FOR DCD-2

Classifier	KERNEL	C	Gamma
Undamaged Car	Linear	1	-
Damaged Texture	RBF	1000000	1e-06
Wheel	RBF	10	10
Headlight	RBF	10	1
Hood	RBF	1	1

- kernel = linear
- C= 1

After tuning the parameters of SVM for best f1 score for each of the classifiers, the best parameters found are given in Table I for DCD-1 and in Table II for DCD-2. For all 5 classifiers, hard negative mining [26] was employed in order to improve the performance.

C. Experimental Results And Analysis

The precision, recall and accuracy for the five classifiers used and the overall system is given in Table III for DCD-1 and in Table IV for DCD-2. The five classifiers were tested on the images that are received from the preceding classifier. The precision-recall curves comparing the performance for the 5 classifiers when tested on both Datasets is given in Fig. 4. It can be seen that the performance of the five classifiers trained and tested for DCD-1 performs better than the five classifiers trained and tested for DCD-2. The reason for this is that DCD-1 consists of images captured at less than 2m from the objects and is of better quality whereas DCD-2 consists of images captured at less than 20m from the objects and is of poorer quality. For damaged texture detection in the second stage both LPB and GLCM were tried and since LBP had a precision of 40%, recall of 38.89% and accuracy of 37%, GLCM with a higher performance as seen in Table III and Table IV was used as the feature extracted in this stage for both Datasets. The overall system accuracy for DCD-1 was 81.83% which is greater than the 64.37% accuracy that was achieved for DCD-2.

TABLE III. PERFORMANCE RESULTS OF THE FIVE CLASSIFIERS AND THE OVERALL SYSTEM TESTED FOR DCD-1

Classifier	PRECISION	Recall	Accuracy
Undamaged Car	67.39	79.49	70.89
Damaged Texture	75.34	83.33	70.70
Wheel	64.03	74.21	71.56
Headlight	64.67	70.45	71.55
Hood	64.99	72.56	70.45
Overall System	60.36	82	64.37

TABLE IV. PERFORMANCE RESULTS OF THE FIVE CLASSIFIERS AND THE OVERALL SYSTEM TESTED FOR DCD-2

Classifier	PRECISION	Recall	Accuracy
Undamaged Car	83.67	83.67	83.67
Damaged Texture	88.03	99.60	88.29
Wheel	80.60	87.87	85.56
Headlight	80.51	87.88	82.6
Hood	79.2	86.2	81.5
Overall System	80	83.75	81.83

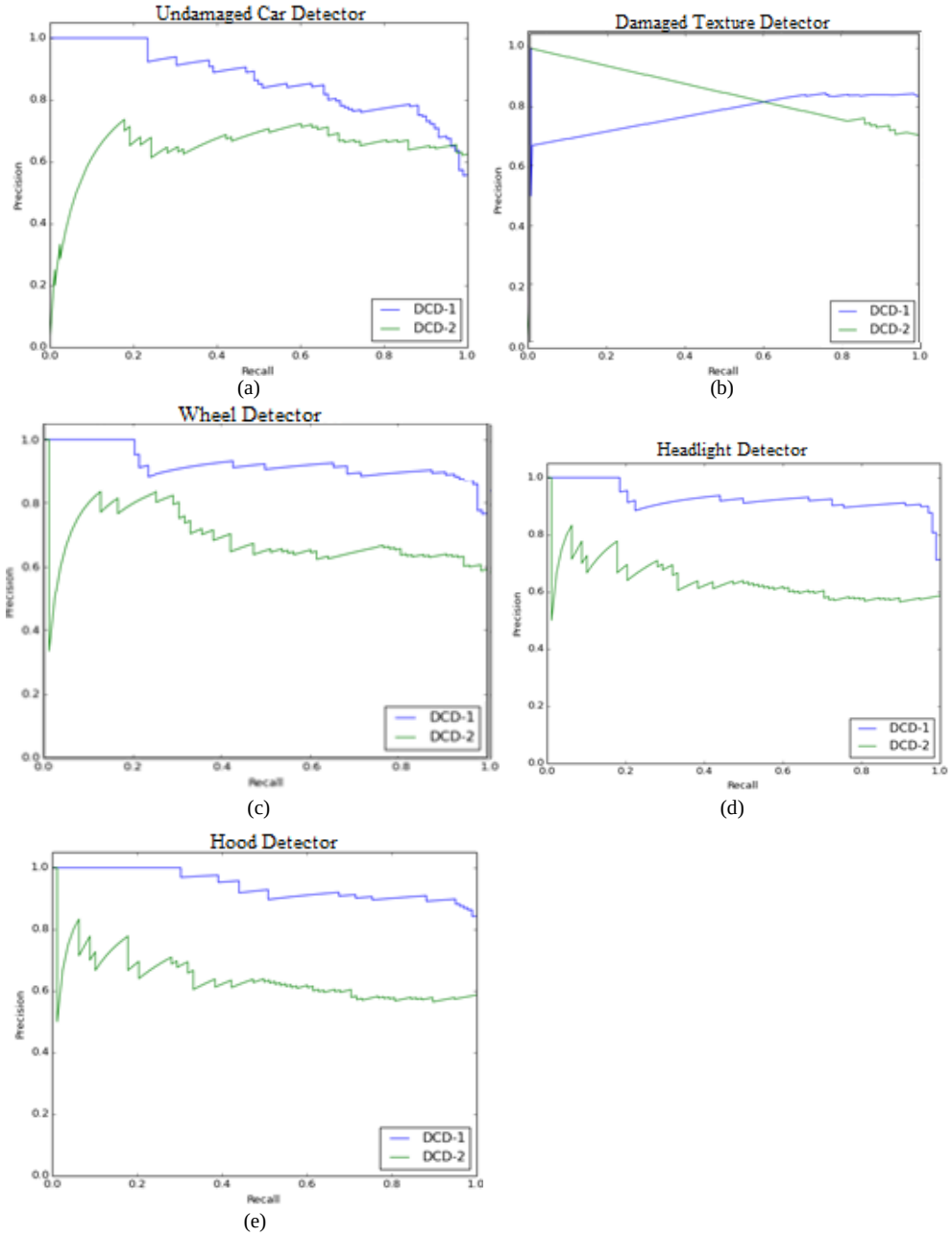


Fig. 4. Precision-recall curves obtained by testing each of the five classifiers' on datasets DCD-1 and DCD-2. (a), (b), (c), (d) and (e) show the precision-recall curves of the undamaged car detector, damaged texture detector, wheel detector, headlight detector and hood detector respectively

V. CONCLUSION AND FUTURE WORK

A new approach to accident detection by detecting damaged cars from footage captured from surveillance cameras has been presented in this paper. Detection of damaged cars using the techniques of Machine Vision was

achieved successfully. The detection was done based on the fact that an undamaged car detector will not detect damaged cars and that most damaged cars have car parts in tact. The method developed used a total of five SVM classifiers trained with HOG and GLCM features.

Since damaged car detection has not been attempted before two datasets of damaged cars were compiled for the sake of the project and released for public use. The system implemented was tested on these two different datasets DCD-1 and DCD-2, which differ based on the distance of the camera from the damaged car, the quality of the images and the number of objects in the images. The accuracy of the system is 81.83% for DCD-1 and 64.37% for DCD-2. However, the system does not detect damaged cars that are damaged to an extant where none of the car parts being considered are present. This is the major limitation of the project.

The future work includes extending the working of the system to detect all types of damaged vehicles as it currently successfully detects only damaged cars. The working of the system can also be extended to detection in nighttime conditions.

REFERENCES

[1] X. Wen, L. Shao, W. Fang, and Y. Xue, "Efficient feature selection and classification for vehicle detection," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 25, no. 3, pp. 508-517, Mar. 2015.

[2] Yichuan Tang. Deep learning using linear support vector machines. In *Workshop on Representation Learning, ICML*, 2013, Atlanta, USA, 2013.

[3] M. Egmont-Petersen, D. de Ridder, and H. Handels, "Image processing with neural networks- A review," *Pattern recognition*, vol. *35, no.* 10, pp. 2279-2301, Oct. 2002.

[4] P.A. Viola and M.J. Jones, "Rapid object detection using a boosted cascade of simple features," in *Proc. CVPR*, 2001, no. 1, pp. 511–518.

[5] "Global status report on road safety 2015", *World Health Organization*, 2016. [Online]. Available: http://www.who.int/violence_injury_prevention/road_safety_status/2015/en/. Accessed: 22- Mar- 2016

[6] V. Goud, "Vehicle accident automatic detection and remote alarm device," *International Journal of Reconfigurable and Embedded Systems (IJRES)*, vol. 1, no. 2, Jul. 2012.

[7] B. Prachi, D. Kasturi and C. Priyanka, "Intelligent accident-detection and ambulance-rescue system," *PULSE*, vol. 450, no. 16, pp. 2, Jun. 2014.

[8] G. Liang, "Automatic traffic accident detection based on the Internet of Things and Support Vector Machine", *International Journal of Smart Home (IJSH)*, vol. 9, no. 4, pp. 97-106, Apr. 2015.

[9] C.Vidya Lakshmi, J.R.Balakrishnan, "Automatic Accident Detection via Embedded GSM message interface with Sensor Technology," *International Journal of Scientific and Research Publications*, vol. 2, no. 4, Apr 2012.

[10] X. Tong, Z. Hong, S. Liu, X. Zhang, H. Xie, Z. Li, S. Yang, W. Wang and F. Bao, "Building-damage detection using pre-and post-seismic high-resolution satellite stereo imagery: a case study of the May 2008 Wenchuan earthquake," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 68, pp.13-27, Mar. 2012.

[11] H. Ma, N. Lu, L. Ge, Q. Li, Z. You, X. Li, "Automatic road damage detection using high-resolution satellite images and road maps," in *Proceedings of the 2013 IEEE International conference on Geoscience and Remote Sensing Symposium*, Melbourne, Australia, Jul. 2013.

[12] S. Sivaraman and M.M Trivedi, "A review of recent developments in vision-based vehicle detection," presented at the Intelligent Vehicles Symposium, 2013

[13] M. C. Motwani, M. C. Gadiya, R. C. Motwani, and F. C. Harris, "Survey of image denoising techniques," in *Proceedings of Global Signal Processing Expo and Conference (GSPx '04),* Santa Clara, Calif, USA, Sept. 2004.

[14] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, 2005, vol. 1, pp. 886-893.

[15] C. Cortes and V. Vapnik, "Support-vector network," *Machine Learning*, vol. 20, no.3, pp. 273–297, Sept. 1995.

[16] C. Hsu, C. C. Chang, and C. J. Lin, "A practical guide to support vector classification," Dep. Comp. Sci., National Taiwan Univ., Taiwan, Tech. Rep., 2005.

[17] T. Ahonen, A. Hadid, and M.Pietik¨ainen, "Face Description with Local Binary Patterns: Application to Face Recognition," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 28, no. 12, pp. 2037-2041, 2006.

[18] T. M¨aenp¨a¨a. The Local Binary Pattern Approach to Texture Analysis — Extensions and Applications. PhD thesis, University of Oulu, 2003.

[19] R. M. Haralick, K. Shanmugam, and I. Dinstein, "Texture features for image classification," in *IEEE Trans. Systems Man Cybemet*, 1973, pp. 610-621.

[20] "ImageNet," [Online]. Available: http://imagenet.stanford.edu. Accessed: 20- Jun- 2016.

[21] "Caltech101," [Online]. Available: http://www.vision.caltech.edu/Image_Datasets/Caltech101/. Accessed: 20- Jun- 2016.

[22] R. Fisher, "CVonline: The Evolving, Distributed, Non-Proprietary, On-Line Compendium of Computer Vision," [Online]. Available: http://homepages.inf.ed.ac.uk/rbf/CVonline/. Accessed: 20- Jun- 2016.

[23] "CIREN The Nation's Largest Learning Laboratory," [Online]. Available: http://www-nass.nhtsa.dot.gov/nass/ciren/SearchForm.aspx. Accessed: 20- Jun- 2016.

[24] "Google-Images," [Online]. Available: https://images.google.com/. Accessed: 18- Apr- 2016.

[25] F. Suard, A. Rakotomamonjy, A. Bensrhair and A. Broggi, "Pedestrian detection using infrared images and histograms of oriented gradients," presented at the IEEE Intelligent Vehicles Symposium, 2006.

[26] T. Malisiewicz, A. Gupta, and A. Efros, "Ensemble of Exemplar-SVMs for Object Detection and Beyond," in *International Conference on Computer Vision*, 2011, pp. 89-96.

[27] "vaishnavi29/DCD", *GitHub*, 2016. [Online]. Available: https://github.com/vaishnavi29/DCD/tree/master/DCD-1/01. Accessed: 10- Jul- 2016.

[28] "vaishnavi29/DCD", *GitHub*, 2016. [Online]. Available: https://github.com/vaishnavi29/DCD/tree/master/DCD-1/02. Accessed: 10- Jul- 2016.

[29] "vaishnavi29/DCD", *GitHub*, 2016. [Online]. Available: https://github.com/vaishnavi29/DCD/tree/master/DCD-1/03. Accessed: 10- Jul- 2016.

[30] "vaishnavi29/DCD", *GitHub*, 2016. [Online]. Available: https://github.com/vaishnavi29/DCD/tree/master/DCD-2/01. Accessed: 10- Jul- 2016.

[31] "vaishnavi29/DCD", *GitHub*, 2016. [Online]. Available: https://github.com/vaishnavi29/DCD/tree/master/DCD-2/02. Accessed: 10- Jul- 2016.

Real-Time Implementation of an Open-Circuit Dc-Bus Capacitor Fault Diagnosis Method for a Three-Level NPC Rectifier

Fatma Ezzahra LAHOUAR[1,2], Mahmoud HAMOUDA[1,3], Jaleleddine BEN HADJ SLAMA[1]
[1]Research Unit SAGE-Research Laboratory LATIS, ENISO, University of Sousse, Sousse, Tunisia
[2]PVT Tunisia Company, 59 Jamel Abdennacer Street, H-Sousse 4011, Tunisia
[3]Canada Research Chair in Electric Energy Conversion and Power Electronics CRC-EECPE, ETS de Montréal, Canada H3C1K3

***Abstract*—The main goal of this paper is to detect the open-circuit fault of the electrolytic capacitors usually used in the dc-bus of a three phase/level NPC active rectifier. This phenomenon causes unavoidable overvoltage across the dc-bus leading therefore to the destruction of the converter's power semiconductors. The real-time detection of this fault is therefore vital to avoid severe damage as well as wasted repair time. The proposed diagnosis method is based on the measurement of the voltages across the two dc-bus capacitors. Their mean values are therefore compared with the half value of the dc-bus reference voltage. If the comparison result is under a predefined threshold value, a fault alarm signal is generated in real-time by the monitoring system. The converter's control algorithm and fault detection method are both implemented in real-time on a DSP controller. The obtained experimental results confirm the effectiveness of the proposed diagnosis technique. Indeed, a fault signal is generated at the peripheral of the DSP after 60 ms of the fault occurrence.**

***Keywords*—fault detection; capacitor failure; open-circuit fault; real-time implementation; multilevel converters**

I. INTRODUCTION

Multilevel converters are without a doubt the best suited topologies for medium voltage/ medium power industrial applications [1]. They have many attractive advantages as compared with two-level converters such as reduced dV/dt, lower voltage stress across power semiconductors, better quality of the output voltages, reduced filter size, etc. [2,3]. Among several power conversion structures, the three-phase three-level Neutral Point Clamped (NPC) is widely preferred since it needs only one dc source and the minimum number of dc-bus electrolytic capacitors.

As any power conversion device, the electrolytic capacitors used in the dc-bus, the power transistors and diodes are the most vulnerable power components of the NPC converter. Usually, the resulting faults remain a serious trouble to be identified in order to avoid significant damage of the remaining converter's components and circuits [4].

Many diagnosis methods have been proposed in the recent literature to identify the open-circuit fault of power transistors and diodes utilized in the three-level NPC converter [5,6,7,8,9,10]. Some of the diagnosis and faulty components identification methods are based on the analysis of the instantaneous pole voltage and its duration [5], as well as the currents in the dc-bus neutral point. The latters (pole voltages and current in the mid-point) are compared with their estimated values which are computed in terms of the switching states and dc-bus voltage or phase currents. Some other methods are based on the analysis of the average current Park's vector [7], the normalized average current [8], the phase current distortion [9], the slope method [10], etc.

The diagnosis of the dc-bus capacitor failures for NPC converters has not been addressed until now. The research works on this issue are restricted right now to the case of three-phase two-level PWM converters, DC-DC converters, and diode rectifiers. In [11], an on-line estimation of the internal Equivalent Series Resistance (ESR) of the dc-bus electrolytic capacitor of a three-phase two-level AC/DC converter was proposed. The method is based on the injection of a ripple term in the phase currents and the analysis of its effect on the dc-bus voltage waveform. The least square algorithm is used to estimate the ESR value of the capacitor. However, this method can be utilized only when the converter is controlled with the space vector modulation scheme (SVPWM). In [12], a capacitor failure detection method was proposed for a three-phase diode rectifier. The method is based on the analysis of the 6^{th} harmonic component of the dc-bus voltage to detect the capacitor aging and open-circuit faults. Though this method is very simple, but it requires the knowledge of the load and supply conditions.

This paper is focused on the diagnosis of open-circuit faults of the dc-bus capacitors in three phase/level NPC converter operating as an active ac/dc rectifier. The detection method is performed by means of the measurement of the average value of the voltage across the upper and the lower dc-bus capacitors. If the comparison result exceeds a predefined threshold value, a fault alarm is generated in real-time. In this manner, harmful effects on other components are avoided and maintenance rescue can be done in safety. Notice that, no additional devices or sensors are needed to implement this method.

The manuscript is organized as follows: Section II presents the three phase/level NPC active rectifier operation principle with a brief description of its electric circuit as well as its current control algorithm. The effect of the open-circuit fault of a dc-bus capacitor on the performance of the converter and the proposed diagnosis method are explained in section III. Experimental results are given in section IV to evaluate the

performance of the proposed fault detection method. Finally, the work is concluded in section V.

II. Operation of Three Phase/ Level NPC Rectifier

A. NPC topology

The three phase/level NPC converter shown in Fig. 1 consists of three legs (k = 1, 2, 3). Each leg is built using four-quadrant switching devices labeled $T_{ki\ (i\ =1,..,4)}$, and two clamping diodes $D_{kj\ (j\ =1,..,6)}$. Moreover, the dc-bus is made with two capacitors connected in series: C_{dc}^{up} is the upper capacitor connected to the dc-bus positive rail; C_{dc}^{low} is the lower capacitor connected to the negative rail. The point of common connection is named midpoint "O". Basically, the two capacitors C_{dc}^{up} and C_{dc}^{low} should have the same value C_{dc}. Accordingly, the voltage across each capacitor is given as follows:

$$\begin{cases} V_{po} = \dfrac{1}{C_{dc}} \int i_c^+ dt + \dfrac{V_{dc}}{2} \\ V_{on} = -\dfrac{1}{C_{dc}} \int i_c^- dt + \dfrac{V_{dc}}{2} \end{cases} \quad (1)$$

For each leg there exist three possible switching states of the power semiconductors designated by "P", "O", and "N". The corresponding pole voltage referred to the midpoint "O" can therefore take three levels: $+V_{dc}/2$, 0, and $-V_{dc}/2$. More details on the operation principle of this power conversion topology can be found in [13].

B. Operation principle as an active rectifier and control algorithm

As illustrated in Fig. 1, in this operation mode, the ac terminals of the converter are connected to the grid voltages through three inductors which operate as low-pass filter to eliminate the high frequency components of the line currents. The dc terminals are connected either to a passive load that consumes the active power or to an active load that provides the active power such as photovoltaic generator or wind turbine [14].

When this converter operates as an active rectifier, two objectives need to be achieved. First, it should regulate the voltage across the dc-bus according to a target reference. Second, it should generate a high quality of the line currents with a perfect control of the reactive power exchanged with the grid. The latter needs to be equal to zero so as to achieve unity input power factor operation. In the upper part of Fig. 1, we illustrate a simplified schematic block diagram of a conventional Voltage Oriented Control (VOC) algorithm that is used to achieve the aforementioned performance. First a PI controller compares the target reference (V_{dc}^{ref}) and the measured value of the dc-bus voltage and computes the reference of the d-axis component of the line current (i_d^{ref}). As for the reference for the q-axis component of the same current (i_q^{ref}), it is set to zero.

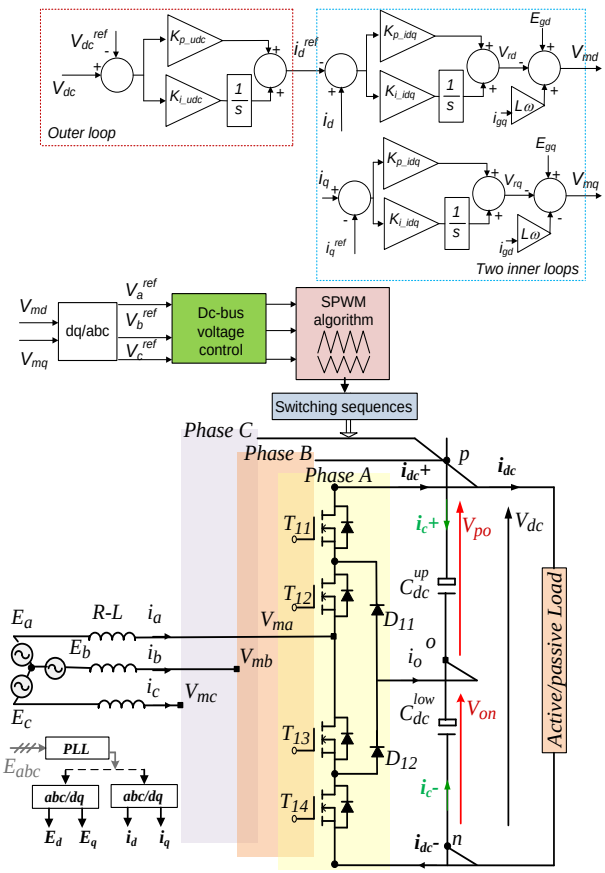


Fig. 1. Grid connected NPC rectifier with its VOC control algorithm

Two additional PI controllers are therefore used to compare i_d^{ref} and i_q^{ref} with the d-q axis components of the measured line currents. The obtained control-laws V_{md} and V_{mq} are transformed into the three-phase reference frame to obtain the suitable modulation signals namely V_a^{ref}, V_b^{ref}, and V_c^{ref}. Finally, the conventional multicarrier PWM algorithm is utilized to generate the appropriate gates pulses of the power transistors.

III. Open-Circuit Fault of the Dc-Bus Capacitor

Basically, the electrolytic capacitors failures are classified into two types [12]. 1) Capacitor aging due to the evaporation of the electrolyte liquid. This phenomenon leads to the increase of the internal Equivalent Series Resistance (ESR) and the decrease of the capacitance as well. 2) The open-circuit fault due to the sudden disconnection of one or more capacitors from the dc-bus. This phenomenon may be due to PCB or connector failures. As a consequence of this abrupt change, a sever overvoltage due to the commutation process of the power semiconductors appears across the dc-bus. Moreover, and inherent unbalance arises between the two voltages V_{po} and V_{on}.

Pasri-Mobidoc/ PVT Tunisia Company

A. *Diagnosis of the open-circuit fault effect on the converter's performance*

To evaluate the effect of open-circuit fault of one dc-bus capacitor on the performance of the NPC rectifier, computer simulations are carried out using a numerical model of the converter. The du-bus voltage reference is set to 100 V. the maximum amplitude of the line-to-line grid voltages is equal to 50 V. The remaining parameters of the power conversion circuit and controllers' gains are reported in Table I. An abrupt open-circuit fault of C_{dc}^{up} was programmed to occur at time t = 1.5 s (Fig. 2). Fig. 3 displays the waveforms of V_{po}, V_{on} and the midpoint voltage $V_o = V_{po} - V_{on}$. One can observe that V_{po} decreases abruptly to zero, after that, its waveform becomes distorted with a chattering phenomenon leading to an overvoltage across the capacitor C_{dc}^{up}. This overvoltage may lead to the destruction of the remaining power semi-conductors. In practice, the measurement and computation of the average value of V_{po} is not easy to accomplish due to the high frequency chattering. As for V_{no}, its amplitude bends down slowly without any distortion or chattering. Therefore, it can be concluded that this voltage contains useful information on the state of health of C_{dc}^{up}. On the other hand, one can also observe that the voltage V_o is affected by this fault where a high frequency fluctuation occurs around the zero voltage value. Fig. 4 illustrates the waveforms of the line currents i_a, i_b, and i_c. One can observe that the waveforms are quite sinusoidal and balanced before the fault. However, their amplitude decreases suddenly after the fault; low-order harmonic components appear also in the line currents waveforms. Therefore, it can be concluded, that this phenomenon causes a substantial decrease of the active power being transferred to the load as well as a distortion of its waveform. Fig. 5-a shows that the phase a line current is in phase with its corresponding grid voltage E_a. However, after the fault, the converter loses its performance of producing a line current in phase with the grid voltage. Fig. 5.b displays the waveform of the q-axis component of the line current. Before the fault, this current is quite equal to zero which means that zero reactive power is exchanged with the grid. However, after the fault, the q-axis current oscillates at low frequency near the zero value. Therefore, the reactive power remains near zero; however its waveform will include low frequency distortions.

TABLE I. EXPERIMENTAL SETUP'S PARAMETERS

Parameters	*labels*	*Values*
Grid frequency	$F_{central}$ (Hz)	50
Switching frequency	f_{sw}(kHz)	3.33
DC bus capacitor	C_{dc} (F)	940e-6
Input filter inductance	L (H)	10e-3
Filter resistor	R (Ω)	1
R-Load	Rload (Ω)	150
Cascaded linear PI controller		
Proportional gain of PI linear parameter of voltage loop V_{dc}	K_p^{Vdc}	5
Integral gain of PI linear parameter of voltage loop V_{dc}	K_i^{Vdc}	10
Proportional gain of PI linear parameter of current controller	K_p^{I}	10
Integral gain of PI linear parameter of current controller	K_i^{I}	250

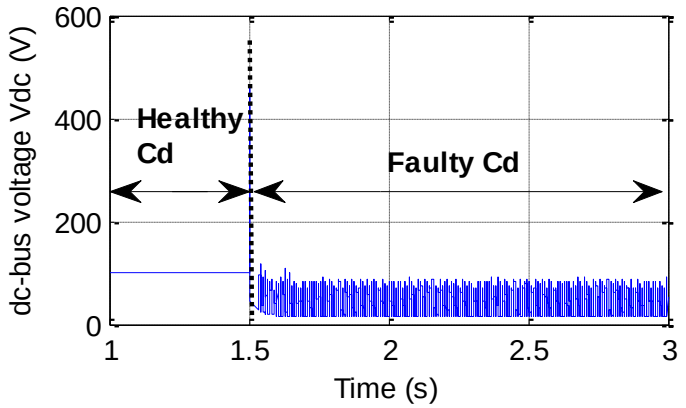


Fig. 2. Dc-bus voltage (Vdc) waveform under healthy and open-circuit faulty conditions (an open-circuit fault of C_{dc}^{up} is triggered at time t =1.5 s)

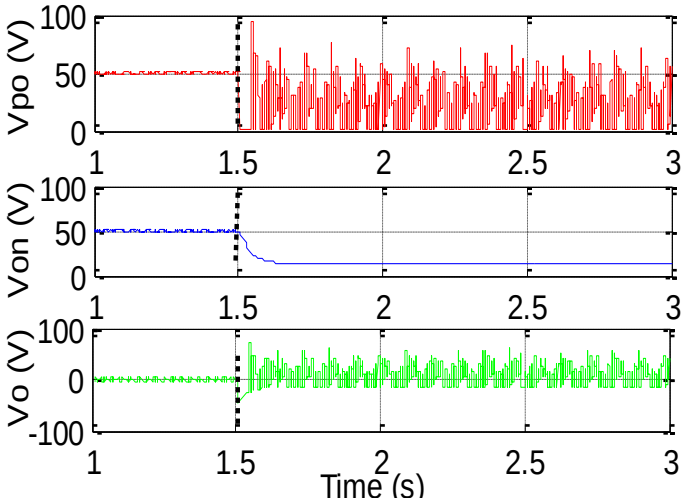


Fig. 3. From top to bottom: Voltages Vpo, Von, and Vo under healthy and open-circuit faulty conditions (an open-circuit fault of C_{dc}^{up} is triggered at time t =1.5 s)

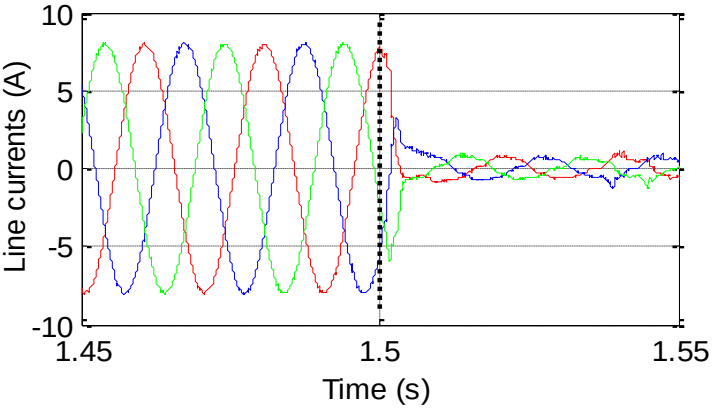


Fig. 4. Line currents waveforms under healthy and open-circuit faulty conditions (fault occurred at t = 1.5 s)

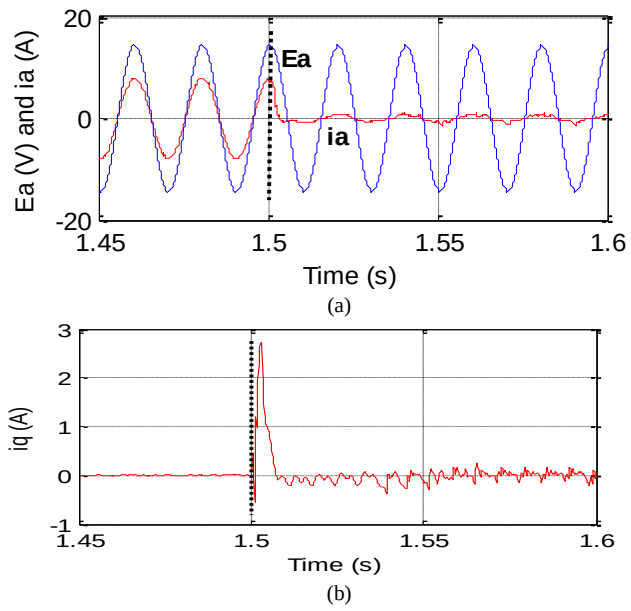


Fig. 5. (a) Line current i_a and line voltage E_a, (b) reactive current i_q under healthy and faulty conditions

B. Capacitor open-circuit fault detection method

Occurrence of open-circuit fault can be diagnosed by monitoring the two dc-bus voltages V_{po} and V_{on}. At each sampling period, these voltages are measured using the same voltage sensors installed with the converter to perform the control algorithm of the rectifier. After that, the average values V_{po}^{Av} and V_{on}^{Av} are computed and compared to the half value of V_{dc}^{ref}. The three following states can therefore occur:

- If $\frac{V_{po}^{Av}}{V_{dc}^{ref}/2} \leq$ predefined threshod value , therefore an open-circuit fault affected the lower capacitor C_{dc}^{low}
- If $\frac{V_{on}^{Av}}{V_{dc}^{ref}/2} \leq$ predefined threshod value , therefore an open-circuit fault affected the upper capacitor C_{dc}^{up}
- Else, both capacitors are healthy.

Fig. 6 displays the flowchart of the proposed diagnosis method.

IV. Experimental Validation based on Real-Time Diagnosis

In order to testify the effectiveness of the proposed open-circuit fault detection method, the VOC algorithm of the NPC rectifier, and the diagnosis algorithm are implemented in real-time on the DSP TMS 320F28335 of Texas Instruments running at a clock frequency of 150 kHz. Both algorithms are edited using Simulink software. The code is thereafter compiled using the real-time workshop of Matlab and Code Composer Studio (CSS). On the other hand, a laboratory prototype of the NPC converter was built using the power transistors IRF460 and the clamping diodes 15ETH06. A twelve-bit analog to digital converter (ADC) with a conversion rate of 80 ns is used for data acquisition. The open-circuit fault is emulated using a solid state relay that is connected in series with the upper dc-bus capacitor C_{dc}^{up}. Fig. 7 illustrates a photo of the experimental setup. The remaining parameters of the experimental setup and controller are quite similar to those used in simulation.

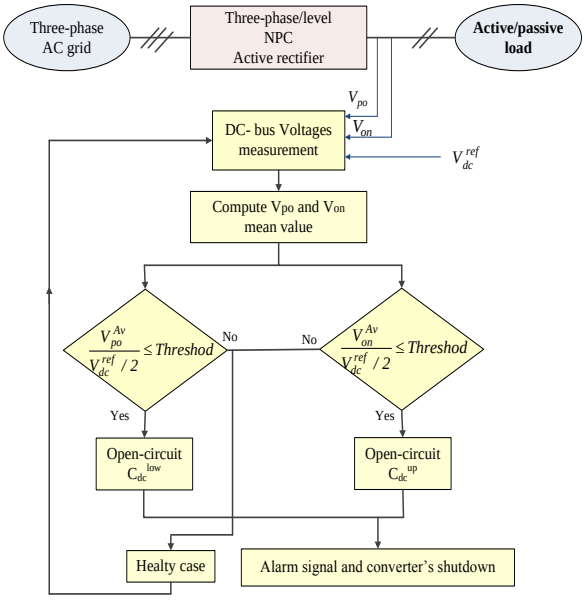


Fig. 6. Flowchart of the proposed dc-bus capacitor fault detection

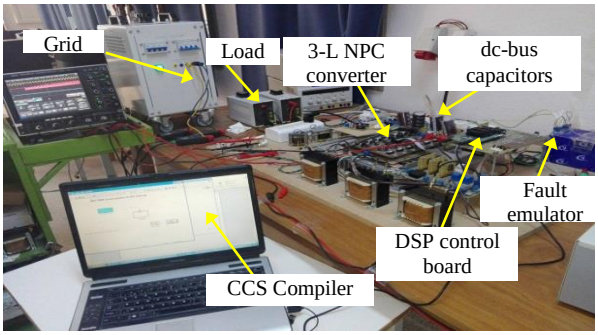


Fig. 7. Experimental setup of a three phase/level NPC active rectifier with faulty dc-bus capacitor

Fig. 8 shows the experimental results obtained under healthy condition and also under an open-circuit fault of C_{dc}^{up}. It is clear that the waveforms of the line current i_a, dc-bus capacitor voltages across C_{dc}^{up} and C_{dc}^{low} , and the voltage V_o are quite similar to those obtained with computer simulations which validates the analysis done in section III. On the other hand, one can observe that the alarm signal is activated after 60 ms of the grid fault occurrence. This is due to the time needed by the voltage across the dc-bus capacitor to decrease under the specified threshold value. The latter was set to 0.9 i.e. $\frac{V_{on}^{Av}}{V_{dc}^{ref}/2} \leq 0.9$. This result emphasizes therefore the effectiveness of the proposed fault-detection method which is very simple and also insensitive to the chattering phenomenon

that affects the waveforms of the voltage across the faulty capacitor.

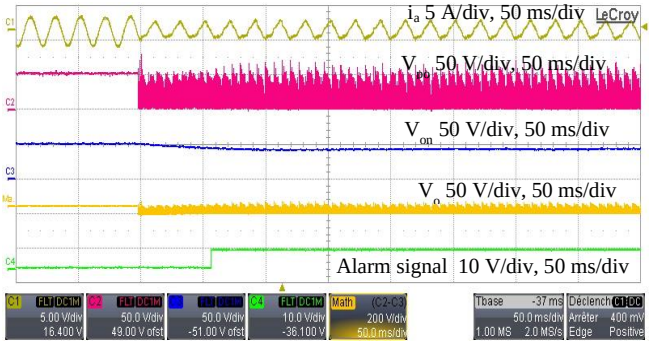


Fig. 8. From top to bottom: line current (ia), Vpo, Von, Vo, and alarm signal

V. Conclusion

A simple and efficient method was proposed with the aim to detect the open-circuit fault of the two capacitors in the dc-bus of a three-level NPC rectifier. The method divides the average value of voltage across each capacitor by the one of the dc-bus reference voltage. If for example the result related to the voltage across the lower capacitor decreases under a specified threshold value, therefore we can conclude that the fault affected the upper capacitor and vice-versa. The algorithm was implemented in real-time on a DSP controller and validated by experimental tests carried out on a laboratory prototype of the converter.

Acknowledgment

This work is carried out in the framework of the "Pasri-Mobidoc" project, and with the collaboration of the industrial partner "Photovoltaïk Technik Tunisia Company". This project is financed by the European Union and administered by the "ANPR".

References

[1] S. Kourou, M. Malinowski, K. Gupakumar, J. Pou, L.G. Franquelo, B. Wu, J. rodriguez, M.A. Rodriguez, and J.I. Leon, "Recent advances and industrial applications of multilevel converters," *IEEE. Trans. Ind. Electron*. vol. 57, no.8, pp. 2553–2580, August 2010.

[2] F. E. Lahouar, J. Ben Hadj Slama, M. Hamouda and F. Ben Mustapha, "Comparative study between two and three-level topologies of grid connected photovoltaic converters," *2014 5th International Renewable Energy Congress (IREC)*, Hammamet, 2014, pp. 1-6.

[3] I. Bouyakoub, B. Mazari, A. Djahbar, and Omar Maarouf, "Simulation of shunt active power filter controlled by SPWM connected to a photovoltaic generator", (IJACSA) *International Journal of Advanced Computer Science and Applications*, vol. 7, no. 8, pp. 65-71, July 2016.

[4] H. Wang, M. Liserre, and F. Blaabjerg, "Transitioning to Physics-of-Failure as a Reliability Driver in Power Electronics," *IEEE Trans. Emerg. Sel. Topics Power Electron*., vol. 2, no. 1, pp. 97-114, March. 2014.

[5] T.J. Kim, W.C. Lee, and D.S. Hyun, "Detection method for open-circuit fault in Neutral-point clamped inverter systems," *IEEE. Trans. Ind. Electron*, vol. 56, no. 7, pp. 2754–2763, July 2009.

[6] H. Jiangbiao, and N.A.O. Demerdach, "An on-line diagnostic method for open-Circuit switch faults in NPC multilevel converters," *IEEE Transportation Electrification Conference and Expo ITEC 2014*, June 15-18, 2014, Dearborn, MI, USA.

[7] A.M.S. Mendes, M.B. Abadi, S.M.A. Cruz, "Fault diagnostic algorithm for three-level neutral point clamped AC motor drives, based on the average current Park's vector," *IET Power Electron*., vol. 7, no.5, pp. 1127–1137, May 2014.

[8] M. B. Abadi, A. M. S. Mendes and S. M. A. Cruz, "Three-level NPC inverter fault diagnosis by the Average Current Park's Vector approach," *2012 XXth International Conference on Electrical Machines*, Marseille, 2012, pp. 1893-1898.

[9] U.M. Choi, J.S. Lee, F. Blaabjerg, and K.B. Lee " Open-circuit fault diagnosis and fault-tolerant control for a grid-connected NPC inverter," *IEEE Trans. Power Electron*., vol. 31, no. 10, pp. 7234-7247, October 2016.

[10] J. S. Lee, K.B. Lee, and F. Blaabjerg, "Open-Switch Fault Detection Method of a Back-to-Back Converter Using NPC Topology for Wind Turbine Systems," in *IEEE Trans. Ind. Appl.* vol. 51, no. 1, pp. 325-335, Jan-Feb. 2015

[11] X.S. Pu, T.H. Nguyen, D.C. Lee, K.B. Lee, and J.M. Kim, "Faults diagnosis of DC-link capacitors in three-phase AC/DC PWM converters by online estimation of Equivalent Series Resistance," *IEEE. Trans. Ind. Electron*., vol. 60, no.9, pp. 4118-4127, September 2013.

[12] T. Kamel, Y. Biletskiy, C. P. Diduh and L. Chang, "Failure detection of the capacitor bank of the three phase diode rectifier," *2012 25th IEEE Canadian Conference on Electrical and Computer Engineering (CCECE)*, Montreal, QC, 2012, pp. 1-4.

[13] P. Chaturvedi, S. Jain, and P. Agarwal, "Carrier-based neutral point potential regulator with reduced switching losses for three-level diode-clamped inverter," *IEEE. Trans. Ind. Electron*., vol. 61, no. 2, pp. 613–624, February 2014.

[14] F. E. Lahouar, M. Hamouda and J. Ben Hadj Slama, "Design and control of a grid-tied three-phase three-level diode clamped single-stage photovoltaic converter," *2015 Tenth International Conference on Ecological Vehicles and Renewable Energies (EVER)*, Monte Carlo, 2015, pp. 1-7.

Issue Tracking System based on Ontology and Semantic Similarity Computation

Habes Alkhraisat
Department of computer science
Al-Balqa Applied University
AL salt, Jordan

***Abstract*—A computer program is never truly finished; changes are a constant feature of computer program development, there are always something need to be added, redone, or fixed. Therefore, issue-tracking systems are widely used on the system development to keep track of reported issues. This paper proposes a new architecture for automated issue tracking system based on ontology and semantic similarity measure. The proposed architecture integrates several natural languages techniques including vector space model, domain ontology, term-weighting, cosine similarity measure, and synonyms for semantic expansion. The proposed system searches for similar issue templates, which are characteristic of certain fields, and identifies similar issues in an automated way, possible experts and responses are extracted finally. The experimental results demonstrated the accuracy of the new architecture, the experiment result indicates that the accuracy reaches to 94%.**

***Keywords*—issue tracking; ontology; similarity computation; vector space model**

I. INTRODUCTION

Issue tracking systems are implemented as a part of integrated project management system. Software projects rely on issue tracking systems to direct corrective maintenance activity and to guide the maintenance activities of software developers. Users report symptoms of the issue along with related information, that include short or detailed textual descriptions of the issue, product and component that are affected by the issue, and how to reproduce. Developers then verify and fix the reported issues. There are often many reports that are received and thus developers would need to prioritize which reports are more important than others.

Issue tracking systems looks like a natural language information retrieval system that can be queried with natural language and return knowledge. Therefore the use of semantic knowledge in developing issue tracking systems improve its ability to semantically infer the similar issues. With cumulative information about issues collected by the issue tracking system over a period and with integrated semantic techniques, it is possible to build semantic issue tracking system that semantically searches for similar issues and links the knowledge for each issue.

This paper, proposes issue-tracking system with integrated semantic techniques for transforming issues content information into meaningful knowledge. The proposed system searches the knowledge documented to inferred semantically similar issues and recommended developers and the most similar files related to the reported issue.

The motivation of this work, for inference of the knowledge field and for recommendation of experts and files related to the reported issue, knowledge document includes not only the previously reported issues but also object-oriented mapping ontology, programmer-readable annotation, and developer experience. The main contribution of the paper is that, it proposes an issue tracking system that predicts similar issue in aa semantic way, possible experts, and possible program files related to issue.

II. RELATED WORK

"Who Knows about That Bug? Automatic Bug Report Assignment with a Vocabulary-Based Developer Expertise Model" [1] uses source code vocabulary to find the most applicable expert for a given bug tracking item. This approach takes long time to make proper recommendations.

"Expertise Recommender: A Flexible Recommendation System and Architecture" [2] uses the change history of source code. It describes a general recommendation architecture that is grounded in a field study of expertise locating.

"Expert Recommender Systems in Practice: Evaluating Semi-automatic Profile Generation" [3] uses a client program which examines the documents within a folder and subfolder which was selected by the user and sends these examined word statistics to the server and compares it with other statistics.

[1], [2], and [3] have the problem that they are not sufficiently integrated into the task workflow a bug tracking and project management system.

III. SYSTEM ARCHITECTURE

The semantic approach for issue tracking system, proposed in this paper, consists of two main process: frontend process for handling the users' issues and backend system for processing issues. Fig. 1 illustrates the main components for issues tracking system proposed in this paper. Frontend system handles the submitted issue and deal with the issue real-time processing. The backend is the platform for frontend processing, and mainly processes and maintains the issue database.

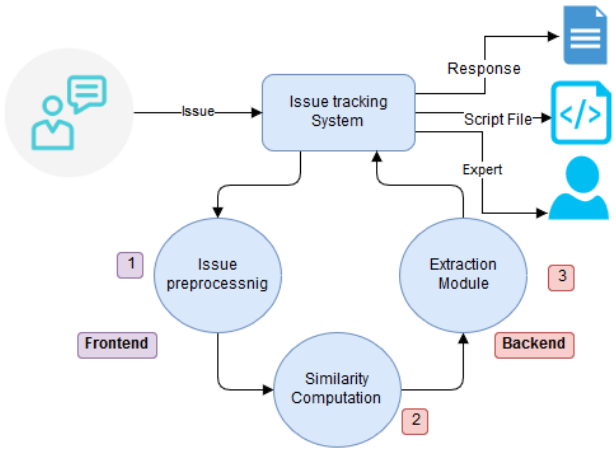


Fig. 1. Overview of issue tracking System Architecture

Fig. 2 shows the global architecture and necessary components to build the proposed system and clarifies the interaction between the system components. The system starts by receiving issue from the user, and finishes by providing the appropriate response for the reported issue, passing through the different phases.

The first component handles the job of prepossessing of issues, source code, and system documentations, which includes tokenization, stemming and keyword extraction. In the second component, the system applies the similarity computation, the similarity computation includes the semantic extension, feature vector generation and similarity measures. In the third component, the response to the submitted issues are extracted from the issues database, for response extraction the following are applied: confidence computation, response selection, and automatic return.

The scenario for an issue tracking system include the following:

1) The user reports an issue to the issue tracking system.

2) Next, in the issue Processing Module, the issue is rephrased by expanding the issue and passing it to the Issue Extraction Module

3) The Information Retrieval component is used to retrieve the relevant issues, response, files, and developers based on the important keywords that appear in the issue.

IV. Preprocessing Model

Preprocessing model starts by tokenizing issues dataset, internal and external system documentations. The tokenization splits up the entire issues, user manual, and programmer annotation into a bag of words. For improving the performance of extracting module and to have exactly matching stems, stemming algorithm has been applied to the bag of words generated after the tokenization process [4]. As a final step, the preprocessing model removes stop words from the bag of words generated after the stemming process. English stop word list which is available online[1] is used for the removal of stop words from the stem word.

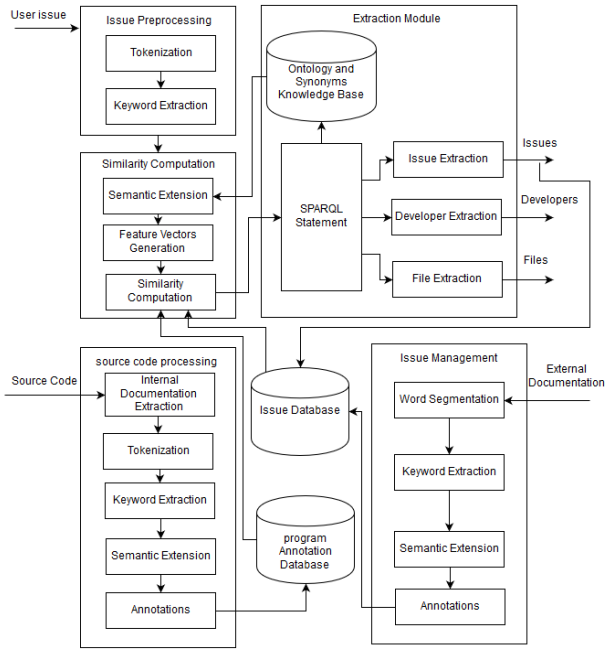


Fig. 2. Global schema of system Architecture

To illustrate the preprocessing model, let's study the following issue:

Issue: There is a problem while uploading the staff personal image.

Tokens: there, is, a, problem, while, uploading, the, staff, personal, images.

Keywords: problem, uploading, staff, personal, image

Stem: problem, upload, staff, person, image

Each keyword is enhanced with the synonyms terms extracted from the ontology. Therefore, our system adds takes the benefits of the shared ontologies and enriches the keyword senses with senses extracted from their synonyms. For semantic extension and keyword enrichment, the synonyms of keywords have been extracted from Macmillan Dictionary[2]. As example, it is possible to enrich the keywords of issue by extraction all synonyms of word "problem" and we get: "problem, difficulty, trouble", and by extraction all synonyms of word "image", and we get: "photo, picture, portrait". The outcome of preprocessing model is a bag of words. The bag of words is then used to represent the issue numerically as vector.

V. Ontology Mapping

Ontologies plays an important role in applications based on the semantic technologies. It consists of concepts, relationships between concepts, restrictions and is described in the ontological languages like Web Ontology Language (OWL).

[1] B. R. Porter M, "The English (Porter2) stemming algorithm," 09 2016. [Online]. Available: http://snowballstem.org.

[2] Princeton University, "WordNet," 09 2016. [Online]. Available: http://www.macmillandictionary.com/.

OWL represents the rich and complex knowledge about things, groups of things, and relations between things.

In some sense, object-oriented representation looks like the ontological representation. In this paper, we used the Semantic framework for mapping object-oriented model to semantic web languages [5]. The relations between object-oriented elements and ontologies are described in table 1.

TABLE I. OBJECT-ORIENTED PARADIGM AND ONTOLOGIES MAPPING

Ontology		Object Oriented
Class	↔	Class
Instance object	↔	Object
Property	↔	Attribute
Predicate	↔	Attribute name
Object	↔	Attribute value

The following example illustrates the mapping process between class written in PHP language and OWL.

PHP class		OWL

class employee
{
private id;
private name;
private personal_image;
}
	→	<owl:Class rdf:id=”staff”>
		<owl:DataTypePropety rdf:id=”id”>
		<rdfs:range rdf:resource=”integer”>
		<rdfs:domain
		rdf:resource=”employee”>
		</owl:DataTypePropety >
		<owl:DataTypePropety
		rdf:id=”name”>
		<rdfs:range rdf:resource=”string”>
		<rdfs:domain
		rdf:resource=”employee”>
		</owl:DataTypePropety >
		<owl:DataTypePropety
		rdf:id=”personal_image”>
		<rdfs:range rdf:resource=”string”>
		<rdfs:domain
		rdf:resource=”employee”>
		</owl:DataTypePropety >

VI. SIMILARITY COMPUTATION

Computing the similarity between user’s issues with both the issues and program annotations databases plays an important role in the automated issue tracking system.

Issue similarity refers to the similarity between the keyword set of given issue annotated by ontology and the pattern keyword set of issues, and instance keywords have been replaced with class keywords in the keyword set of ontology.

There are many computational models for text similarity such as, support vector machines (SVMs), neural network (NN), machine learning, K-Nearest Neighbor (KNN), and so on. In this paper, vector space model (VSM) has been applied for implementing the propose issue tracking system. The VSM was developed for the SMART information retrieval system [6]. VSMs perform well on tasks that involve measuring the similarity of meaning between words, phrases, and documents [7].

The idea of the VSM is to represent large collection of documents as a vector in a vector space. Using VSM the set of issues and queries are represented as m-dimensional vectors of identifiers in a common vector space, and the vectors are organized into a matrix. The row vectors of the matrix correspond to words and the column vectors correspond to issues. Suppose the issue collection contains n issues and m unique terms. The vector space will then have m rows and n columns. The element $w_{i,j}$ in document vector space represents a non-binary weight of the i^{th} term k_i in the j^{th} issue I_j. Let the weight $w_{i,j}$ associated with a pair (k_i, I_j) is positive and non-binary, then the issue I_i and query Q are represented as vectors:

$$\vec{I_i} = (w_{i,1}, w_{i,2}, \cdots, w_{i,m})$$
$$\vec{Q} = (w_{q,1}, w_{q,2}, \cdots, w_{q,m})$$

where m is the number of feature terms.

The relevance of an issue to a query is given by the similarity of their vectors. The weight for terms in queries and issues are used in the computing degree of similarity. The most popular way to measure the similarity of two vectors is to compute their cosine. The cosine of the angle between issue vector $\vec{I_i}$ and query vector $\vec{Q}$ in vector space models can be measured as follows:

$$sim(\vec{I_i}, \vec{Q}) = \cos(\theta) = \frac{\vec{I_i} \cdot \vec{Q}}{\|\vec{I_i}\| \times \|\vec{Q}\|} = \frac{\sum_{i=1}^{t} w_{i,j} \times w_{i,q}}{\sum_{i=1}^{t} w_{i,j}^2 \times \sum_{i=1}^{t} w_{i,q}^2} \quad (1)$$

$sim(\vec{I_i}, \vec{Q})$ varies from 0 to +1, the vector model ranks the documents according to their degree of similarity to the query, the $sim(I_i, Q) = 1$ when $I_i = Q$, and $sim(I_i, Q) = 0$ when I_i shares no terms with Q.

VII. TERM-WEIGHTING SCHEME

The success of vector space model lies in term-weighting scheme. It assigns more weight to surprising events and less weight to expected events. The term weighting for the vector space model has entirely been based on single term statistics. There are three main factors: term frequency factor, collection frequency factor and length normalization factor. All three factor are multiplied together to make the resulting term weight.

In VSM a weight is assigned to each term in a document depends on the number of occurrences of the term in the issue. The frequency of a term k_i inside a document I_j referred to as the term frequency tf factor and is given by:

$$tf(k_i) = \frac{freq_{i,j}}{max_j\, freq_{i,j}} \quad (2)$$

Furthermore, the inverse of the frequency of a term k_i among the documents in the collection referred to as the inverse document frequency idf. idf measures of rareness of a term across all documents. Assume there are N documents in the collection, and that term k_i occurs in df_i of them. Then idf factor of term k_i is essentially

$$idf(k_i) = \log(\frac{N}{df_i}) \quad (3)$$

Thus, the idf of a rare term is high, whereas the idf of a frequent term is likely to be low. The tf and idf are combined, to produce a composite weight schema for each term in each document, the resulted weight schema is called a Term frequency–inverse document frequency $tf - idf$ scheme

[7][8]. The $tf - idf$ weighting scheme assigns to term k_i a weight in document d given by:

$$tf - idf(k_i) = tf(k_i) \times idf(k_i) \quad (4)$$

VIII. Response Extraction and Updating Issues Database

Once ontology classes and properties from keyword set are well fixed, the appropriate SPARQL query to retrieve the response from issue database is built[3]. The query is composed from resources and ontology classes determined by the keyword set. The class name in query pattern will be replaced with corresponding instance name in user issue, based on the query patterns and replace-pair in the most similar issue. The response to the issue is parsed to get extracted the response from the result of SPARQL statement.

Finally, the system updates the issues for constant learning and answering new issues. All issues with a cosine similarity measure higher than a defined cut-off threshold are considered similar and added to the feature set, weighted by its similarity value. On the other hand, if the a cosine similarity measure is less than the predefined threshold, then there is no corresponding query pattern in the issue database, and the issue will be added to issue database after annotating. In this paper, all issues with a cosine similarity value higher than 0.17, which has a 76% accuracy, are considered similar and added to the feature set, weighted by its similarity value.

IX. System Implementation

Issue tracking system architecture described previously has been implemented using PHP. It is semantic-based issue response that returns similar issues and recommends experts for the submitted issue.

X. Experiments and Evaluation

For the evaluation purpose, the proposed issue tracking system has been installed for employees working at IT department of an Institute of Family Health (IFH)[4]. At the time of system evaluation set of 240 issues have been evaluated collected from 100 employees working at IFH and 5 experts working at IT department have been given. The Ontology database includes 20 classes and 100 properties.

For evaluation purpose, accuracy and Recall Rate has been applied [9]. Accuracy and recall for the proposed issue tracking system are shown in Table 2.

[3] https://www.w3.org/TR/rdf-sparql-query/

[4] The Institute for Family Health (IFH) is a regional model providing comprehensive family healthcare services and training for professionals and caretakers in the fields of family healthcare. http://www.ifh-jo.org/index.php?language_id=1

XI. Conclucsion

The main contribution of this paper is that it proposes an ontological semantic based issue tracking system. To achieve the system goals, the proposed system combines the Vector Space Models with domain ontology representing the issue and code vocabulary. To improve the semantic similarity and accuracy of system, each word is enhanced using the synonyms terms extracted from the ontology pool and keywords synonyms database. Therefore, the system takes advantage of the shared ontologies available on the Web and semantically enriches the keyword senses with senses extracted from their synonyms.

The experimental results indicate that the system reaches an accuracy of 94% based on test set of 240 issues and 5 experts. In future extensions, the accuracy of the proposed system would be compared with Jira bug tracking [10].

TABLE II. Experiment Result

	Issue Response	**File recommender**	**Experts recommender**
Recall	95%	98%	92%
Accuracy	94%	96%	93.5%

References

[1] M. Dominique , K. Adrian and N. Oscar, "Assigning Bug Reports using a Vocabulary-Based Expertise Model of Developers," 6th IEEE International Working Conference on Mining Software Repositories, pp. 131-140, 2009.

[2] M. W. David and A. S. Mark, "Expertise Recommender: A Flexible Recommendation System and Architecture," in Proceedings of the 2000 ACM conference on Computer supported cooperative work, New York, 2000.

[3] T. Reichling and V. Wulf, "Expert recommender systems in practice: evaluating semiautomatic profile generation," in SIGCHI Conference on Human Factors in, New York, 2009.

[4] M. Porter, "An Algorithm for Suffix Stripping," Program electronic library and information systems, vol. 14, no. 3, pp. 130-137, July 1980.

[5] M. R. Ježek P., Semantic framework for mapping object-oriented model to semantic web languages., vol. 9, Front. Neuroinform, 2015.

[6] G. Salton, A. Wong and S. C. Yang, "A Vector Space Model for Automatic Indexing," Communications of the ACM, vol. 18, no. 11, pp. 613-620, Nov. 1975.

[7] D. M. Christopher, R. Prabhakar and S. Hinrich, Introduction to Information Retrieval, New York: Cambridge University Press, 2009.

[8] G. Salton and C. Buckley, "Term-weighting approaches in automatic text retrieval," Information Processing and Management, vol. 24, no. 5, pp. 513-523, 1988.

[9] S. Blair, "A Guide to Evaluating a Bug Tracking System.," 2004.

[10] V. Heyn and P. Adrian, "Semantic Jira - Semantic Expert Finder in the Bug Tracking Tool Jira," in 9th International Workshop on Semantic Web Enabled Software Engineering, Berlin, 2013.

Constraints in the IoT: The World in 2020 and Beyond

Asma Haroon
Department of Computer Science
COMSATS Institute of Information Technology
Islamabad, Pakistan

Wajeeha Naeem
Department of Computer Science
COMSATS Institute of Information Technology
Islamabad, Pakistan

Munam Ali Shah
Department of Computer Science
COMSATS Institute of Information Technology
Islamabad, Pakistan

Muhammad Kamran
Department of Distance Continuing & Computer Education,
University of Sindh,
Hyderabad, Pakistan

Yousra Asim
Department of Computer Science
COMSATS Institute of Information Technology
Islamabad, Pakistan

Qaisar Javaid
Department of Computer Science & Software Engineering,
International Islamic University,
Islamabad, Pakistan

***Abstract*—The Internet of Things (IoT), often referred as the future Internet; is a collection of interconnected devices integrated into the world-wide network that covers almost everything and could be available anywhere. IoT is an emerging technology and aims to play an important role in saving money, conserving energy, eliminating gap and better monitoring for intensive management on a routine basis. On the other hand, it is also facing certain design constraints such as technical challenges, social challenges, compromising privacy and performance tradeoffs. This paper surveys major technical limitations that are hindering the successful deployment of the IoT such as standardization, interoperability, networking issues, addressing and sensing issues, power and storage restrictions, privacy and security, etc. This paper categorizes the existing research on the technical constraints that have been published in the recent years. With this categorization, we aim to provide an easy and concise view of the technical aspects of the IoT. Furthermore, we forecast the changes influenced by the IoT. This paper predicts the future and provides an estimation of the world in year 2020 and beyond.**

***Keywords*—Internet of Things; Future Internet; Next generation network issues; World-wide network; 2020**

I. Introduction

The Internet of Things (IoT) is a next generation world-wide network that contains large number of interconnected heterogeneous physical devices, enlightened by [1]. CompTIA's research [2] and M. Swan [3] discussed the estimated ratio of objects connected to the internet to reach 50 billion in 2020. In fact, Cisco estimated the 11 trillion devices per year over 2025 [4]. This high-level connectivity supposed to be delivered by the IoT will clearly play main role in technical advancements, which will open new ways of productivity with more flexibility and customization. Technically, a device in specification of the IoT; is an object in real-time environment implemented with improved capabilities of computation and communication. Theoretically, the IoT is a durable connection of aforesaid objects or devices.

Nevertheless in the way to progress, the IoT faces some technical requirements such as functional requirements, non-functional requirements and design constraints categorized by [5]. Design constraints can be elaborated by the fact that smart things generally regarded as small sized physical devices connected with the Internet, face limitations in terms of IP numbers, packet size, packet loss and alternative paths for connectivity, throughput, power and supported complexity [6]. Furthermore, connected objects vary according to the functionality or purpose they serve. Elmangoush *et al*. [7] explained the heterogeneity and the challenges related to connecting devices that includes scalability, governance and lack of testbeds specifically for Smart Cities. The nature of each object varies according to their size, position and capabilities.

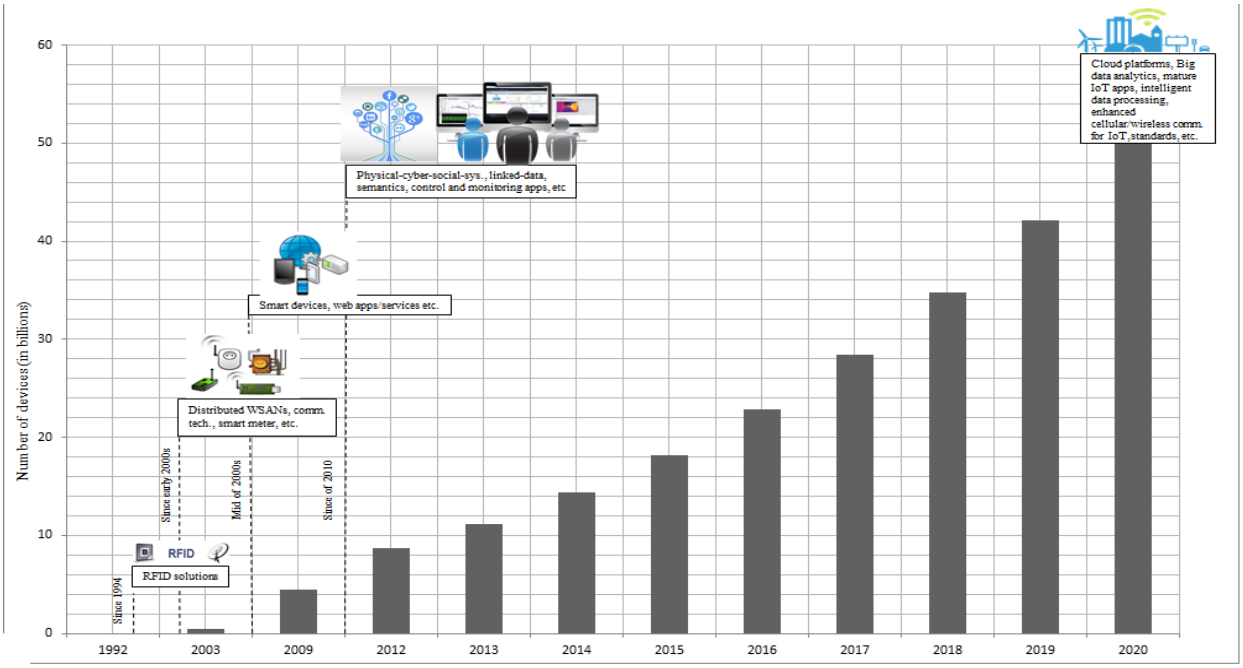


Fig. 1. Estimated number of interconnected devices obtained from [2][13]

In order to connect aforementioned objects in a network, a unique identification number such as RFID (Radio Frequency Identification) and sensor(s) for their continuous state sensing, are needed. RFID and sensors will enable the objects to interact but due to the huge ratio of objects and their variations creates the scalability and diversity of objects; a major hurdle. This will ultimately address the challenge to create a communication framework in such a way that can handle the scalability and variety of objects to achieve the intended applications that may encompass automation functionalities. Automation functionalities include sensing, acting, locating, identifying etc. analyzed by [8]. Figure 1 estimated the compound growth of the IoT that contains both ratios of existing and expected connected devices, along with their diversity. The curve in the Figure 1 also elaborates that the number of physical devices (in billions) per year is adversely increasing, that addresses both standardization and scalability challenge.

The IoT deployment demands to handle the challenges regarding constraint devices, scalability and diversity security issues. That ultimately leads to the requirement of enabling technologies such as hardware, software and algorithms, standardization techniques, network technologies, communication technologies, power and storage technologies, identification technologies, sensing technologies, data transference technologies, architecture and network relationship management technologies addressed by [9][10][1]. The major limitation involved in the deployment of the IoT regarding these technologies are analyzed by [7] [9][11][12] such as micro software, power and storage need, communication standardization, communication protocols, architecture design, security and network management. Multiple numbers of unique IDs per object need leads to the requirement of infinite number of unique IDs. However, this problem has been solved to some extent by IPv6. Micro operating systems are needed for unmeasurably small devices. Power and storage is the major requirement for small devices to sustain the reliable connectivity to the IoT. The necessary standardization techniques are required for the communication between diverse natures of devices in a network. Moreover, the considerable fact is that all of the devices or things are not connected to the Internet, all the time. In that case, there is a need to manage an alternative communication path. Therefore, architecture design also faces the barrier of devices' nature in this regard and checks for an alternative way to be connected. Security concerns and network management are also referred as major hurdles in the IoT.

The basic principle of the IoT technologies has been discussed in [10] is a combination of Sensor technology, RFID technology, Smart technology and Nano-technology. Aforesaid technologies are not new but with IoT advancement these technologies are been more focused and enhanced. However, there are lots of challenges with the addition of technical constraints in the way to deployment of the IoT that will be surveyed in this paper.

Contribution of the paper:

- This paper provides an analysis of recent major technical issues that the IoT is currently facing. The taxonomy for categorization of the IoT is provided.
- Moreover, a categorization of the technical constraints and their overlapping factors is also presented.
- The brief survey for technical restrictions of recent papers is presented in tabular form. It is positioned as a survey paper beneficial for wide range of audience such as business strategists, data analysts, researchers etc.

- Furthermore, the technological advancements in the IoT along with their limitations and effects on society are provided to update the readers with recent trends.
- The open social challenges in future are also highlighted in the paper.
- The paper aimed to refine the basic principles needed to deploy the IoT.
- Moreover, the paper provided the estimation on the ratio of interconnected devices to the IoT up to 2026 and ratio of IT jobs up to 2020 based on previous estimations.

The rest of the paper is organized as follow. In Section II, the IoT technical constraints and enabling technologies are analyzed in broader perspectives. Most recent technological advancements related to the IoT are described briefly in section III to introduce the readers with the technological updates. Section IV is about performance evaluation of the technological advancement with their impacts on society and technical limitations. Whereas, in section V remaining open issues related to the IoT are shortly listed. Finally, to provide a short summary paper is concluded at the end in section VI.

II. Technical Constraints

The Internet has been changed from computer-to-computer connection to ubiquitous Internet and now proceeding towards the IoT; that is everything is interlinked with every other thing, anywhere and anytime. It indicates that the IoT will provide bases to initiate a new technological phase soon (estimated in 2020), which will expose new means of opportunities in everyday life. Due to its vast applications, the IoT has been focused and new ideas have been proposed in this regard by many researchers in recent years. Figure 2 enlightened the major areas of the IoT that includes the IoT challenges, the top IoT enterprises, applications of the IoT services and the IoT solutions. However, the paper aims to specifically focus and explore the further categorization of its technical barriers highlighted by Figure 2. This section will analyze the technical constraints that include standardization and policies issues, hardware limitations, gateway systems challenges, middleware issues, the database management issues and security and privacy challenges as described below.

A. Addressing and Sensing issue

In the IoT, every object in real time environment, either it is a living thing or non-living thing, needed to be addressed by a unique identity. [11][12][19][20][21][22] analyzed the addressing and sensing issues in the IoT perspective such as IPv6 adaptation, automatic identification and configuration, participatory sensing, etc. Using sensor technology networks, it is obvious to have large number of nodes that must be addressable separately. On the other hand, the problem is the ratio of objects is far greater than IPv4 addressing scheme. Future estimations predict that it will increase to infinite number of devices or object instead of decreasing [2][4][13]. B. Stockebrand [23] claimed that IPv4 is already outnumbered and all of the IP addresses had been occupied. Therefore, IPv6 was defined by the means of 128 bit which will fulfill the demands of ever increasing IP addresses.

The famous claim about IPv6 is stated as it can assign address to every bit of this world but this scheme still faces the design limitations. The approach cannot be used in a scenario; if RFID tag identifier is 96 bits long. L. Atzori *et al*. [10], analyzed the methodology for this problem in which a separate agent is used for IPv6 IP address as an interface ID. As identified by [24], the transition to pure IPv6 is pretty challenging but it have more advantages over IPv4 scheme such as providing internal security and end-to-end user transparency with the addition of realization of addressing need. However, transition is not that easy, the IPv6 users' needs IPv4 side by side to use important resources and stay connected to existing connection. IPv6 requires a large amount of time to be fully functional as a standalone scheme. It shows that there is a huge gap for mobility support on technical bases. Therefore, mobility management is needed to be focused because of adaptability and scalability issues related to diverse nature of devices in heterogeneous environment.

The order in which address is been fetched correspondingly faces the drawbacks related to addressing technique. As Domain Name Server (DNS) is used for mapping domain name while fetching the IP address of host associated with specific given name but in the IoT, communication must take place between objects rather than hosts. The issue has been addressed by [20][25][26] and Atzori *et al*. [10] proposed that an Object Name Server (ONS) technique can be used at object level communication. In current Internet, the tag identifier mapped on the Internet Uniform Resource Locator (URL), and the desired information is fetched. However, Object Name Address (ONA) should work in both ways and vice versa in the IoT. It should be able to associate the information to a given RFID tag identifier and can also map on in opposite direction. The other issues in addition to addressing is; if the device is connected to Internet all the time then it can be addressable and its state can be sensed. However, due to heterogeneity of devices in the IoT, all of the devices are not connected to the Internet directly according to their nature or some critical security issues. In that case, the devices need some technique to be addressed indirectly and sensed to get updated by its state through some other medium.

B. Networking Issue

In networking, protocols play a critical role for connection and data transformation; as it can reduce the service integration time and cost [7][17]. Network protocol acts as a mainstay for data routing between outer world and sensors. The current Internet is using TCP protocol for transmission at transport layer, which is not feasible for the IoT due to its limitations. There are a lot of existing protocols depending upon different criteria for mobile networking but all of them have drawbacks making them impracticable for the IoT. So there is a need of protocol for efficient handling and processing of data. [9][10][11][27][28] analyzed the major issues related to TCP protocol which can be categorize as:

1) Connection Setup: TCP protocol creates connection first before initiating any data transmission which means, it is connection setup based protocol. It takes considerable time for

creating connection before actual data transmission. It seems unnecessary and time wastage in case of the IoT because the amount of data and time for connection is considerable short. Furthermore, connection is created between two terminals which is energy and resource taken, so it is not feasible.

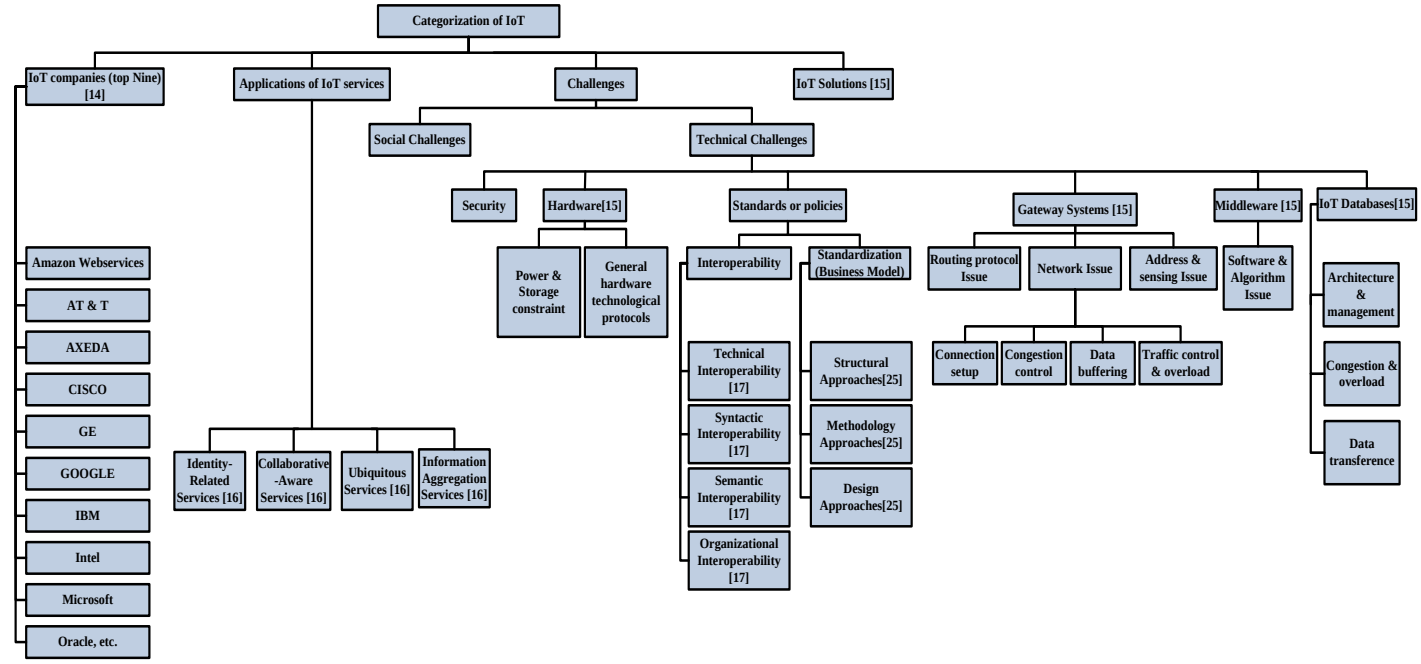


Fig. 2. Categorization of IoT obtained from [14][15][16][17][25]

2) Congestion Control: TCP protocol is responsible for performing congestion control over both terminals during data transmission, which is not realizable in case of the IoT due to its heterogeneous nature. Most of the time, data to be transfer is of small size and the congestion control in that case is an overload. Moreover, the communication is done between different types of wireless networks and mediums. Congestion control in that scenario will decrease the performance. Thus, the TCP congestion control with its existing state is impractical in the IoT perception.

3) Data Buffering: TCP protocol stores data at both terminals to ensure the secure transmission of data. Subsequently, in case of any damage or loss during the transmission data can be resent. It requires buffers on both ends to store the data that will be very costly in term of both energy and storage for the devices which are small with low storage capacity and very limited battery life.

4) Traffic Control and Overload Issue: Traffic control in the IoT is another challenging task related to networking. It is a smooth transmission in term of traffic control when it is only between sensor nodes in wireless network. But it become complicated when sensors become part of whole network having heterogeneous purposes. In machine-to-machine (M2M) communication, the traffic control is totally different than human-to-machine communication. Moreover, unmeasurable number of devices involve in the IoT will also create overload traffic issues. Therefore, there is a need of characterization of network traffic which totally depends upon application scenarios. Furthermore, the existing network infrastructure is unable to address the high amount of traffic that is going to be generated in near future. Efficient protocol having advance levels of traffic handling and network management are needed to be implemented.

C. Routing Protocol Issue

Vehicle-to-Vehicle (V2V) communication is a type of distributed computation environment, which has huge number of nodes with variable and constrained network topology. However focusing on the importance of routing aspect in V2V communication, S. Agrawal and D. Vieira [11] discussed the two basic ways of routing. One is source routing: in which destination is already defined. Second is hop-to-hop routing: in which only next node address is known. Therefore, hop-to-hop routing is more suitable for V2V communication. Thus, the next best hop can be selected for routing during communication.

Routing protocols like Geographical Source Routing (GSR) use global positioning mechanism, which can cause path uncertainty and route fluctuation. On-Demand Routing protocol use flooding method that can create congestion problem because it sends data to all possible nodes. Various other existing routing techniques like Greedy Perimeter Stateless Routing (GPSR), Dynamic MANET on Demand, etc., have their own limitations and drawbacks. M2M is a key enabler for Smart Cities addressed by [7]. With the advancement of technologies, M2M will require consistent data routing due to the need of high data rates. This is a key challenge to create a reliable routing protocol having high speed transmission and low delivery delay time.

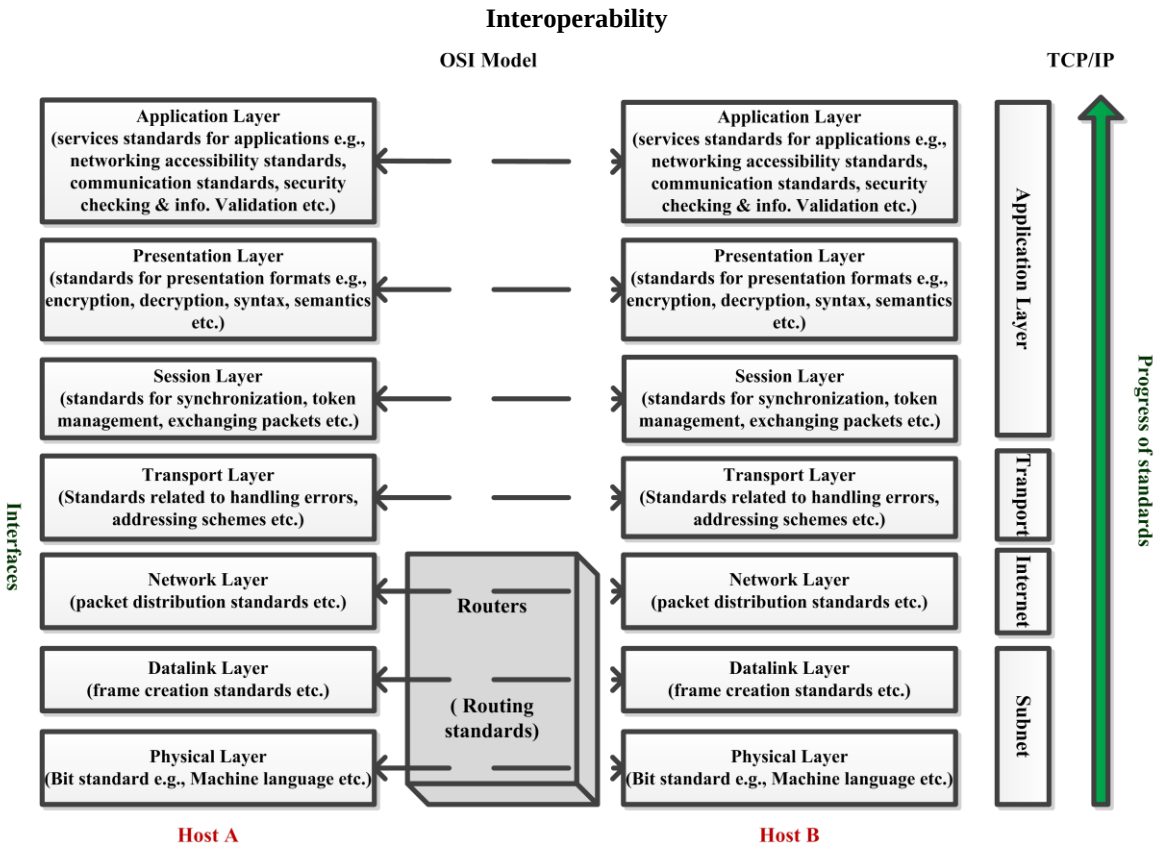


Fig. 3. Standardization level with respect to OSI model within interoperability obtained from [61] [62] [63]

D. Standardization Issue

The number of items in the IoT is extremely high. Therefore, issues related to representing information, storing information, interconnections, searching, and organizing information produced by the IoT will become very challenging [10]. Different approaches categorized as structural approaches, methodology approaches and design approaches have been proposed to achieve the business model discussed by [18]. In current Internet, application domains are separate, which is making business domains separate and ultimately not meeting the goal of the IoT. Thus, there is a need of standardization techniques to join all separated application domains in a sophisticated manner.

Diversity is an inherited characteristic of the IoT that leads to the major issue refer as interoperability. Whereas, standardization of technologies lead to better interoperability; as interoperability depends upon standards for functions and interfaces expressed in Figure 3. The gap that needs to address is lack of technical interoperability among diverse devices. W. Pollard [29] stress that in M2M communication scenario, standardization techniques provide a middleware to handle communication mechanisms, device management and reachability between end terminals. Moreover, it is necessary to follow standards for devices to work together on the same platform of the IoT. Otherwise the devices will not provide proper services to client. International Organization for Standardization (ISO) provide family of standards, which are gaining popularity [17][30].

M2M communication is leading model toward the IoT but there is precise little standardization work done in this regard. The aim of M2M communication is to connect all the devices, sensors and their actuators abstracting communication techniques. The operation cost for M2M communication is also addressable. Therefore, there is a need of optimized standard interfaces to be made for M2M communication in order to address interoperability and scalability issues related IoT. European Telecommunications Standards Institute (ETSI) is focusing on standardization techniques for M2M communication. Separate technical committee is launched for this purpose to increase the effort speed in European Telecommunications Standards Institute (ETSI). The common issues included are location, addressing, sensor networking integration, naming, charging, Quality of service (QoS), privacy and security, network management, software or application, and hardware interfaces for M2M communication standardization [7], [11], [28], [31], [32], [33].

E. Software and Algorithm Issue

There is a need of common software (in terms of new protocols) and algorithms to provide a middleware base independent of resources and networking function for the connectivity in different environments among diverse devices.

J. Gubbi et al. [12] focus on creation of such distributive application development scenario to build a coherent application. That will support interoperable interaction among M2M communication over a network. In [34], it is focused that the distributive application should be containing self-manageable properties containing self-optimization, self-configuration, self-healing to handle communication in different scenarios.

New micro operating systems are also required that can efficiently function for small devices in terms of energy and power. New password mechanisms should also be introduced to ensure the security and privacy during the communication. In this regard Service Oriented Architecture (SOA) approach has been followed by the IoT [10][21]. Service Oriented Architecture (SOA) allows decomposing the complex systems and monolithic system. This results in well-defined and simpler application development that follow standards during development ultimately facilitates components coordination among each other. This approach also gives reusability of software and hardware components. Perhaps, it also misses the solution to abstraction of details including devices, functionalities and capabilities details. To maintain the position on top levels CISCO claims to provide such software solution that will be highly focused on security concerns by [35].

F. Power and Storage Constraint

The IoT device is constrained by the entity which is in physically monitoring state and the position of the entity is frequently changing without access to power [7]. Most of the devices in the IoT are having considerable small size and are not fixed. Due to their size and frequently changing location property devices are not able to access the power all the time. So the low power consumption is universal constraint of the IoT. Either they use battery technologies or they can use some techniques for taking power from their environment using other devices. Therefore, there is a need of design in which such power consumption techniques or low power consumption schemes are made with long lasting life of devices. Another design issue to be addressed is a requirement of such modular approach that subtracts the need to make a separate chip for each and every application because it is not feasible to create a chip for each and every application. Such high modular approach will combine existing chips within size and power constraint. In [25], some theoretical low power circuit solution is proposed. Nevertheless, this area needs a lot of research and effort without which the IoT cannot be achieved.

G. Architecture and Network Relationship Management Issue

In the IoT perspective, devices have not been expected to sustain their positions. However, the reliable connectivity demands to address and sense the devices all the time. To address huge number of mobile nodes in a world-wide network is also referred as major scalability issue. [7][12][10][19][36][37] analyzed the need to address architecture and network relationship management issue. N. Meghanathan, S. Boumerdassi, N. Chaki, and D. Nagamalai [38] emphasized the need to build architecture with such efficient mechanism that can discover all sensor resources and can register and update new sensing systems in wider network. In [39], the major challenge to the IoT is triggered that "*Who monitors the monitor?".* The IoT is integrating the heterogeneous devices into already defined networks, which is particularly advancement for industrial revolution. The problem is; if the information from machines are stolen or fetched and used for purpose by unauthorized person or entity. The IoT will never stop evolving and ratio of objects will increase with the time. With more devices connected to the IoT will raise the amount of data that required to be managed efficiently. Moreover, it is not possible to cover each and every scenario of the IoT. By customizable sensor technologies the things are able to be monitored by a specific location but the things can be monitored not the devices.

Intelligent networks are needed to be created which can control and monitor each other independently. Another issue pointed by CISCO [40], is data collection especially in Smart Grids Networks, which clearly maps to scalability issue. Other issues related to networking management include Closed-Loop Functioning and Network Resource Preservation. There is a need of such architecture which will able to handle these sort of Ad-hoc networks using different Network technologies (wireless, fixed, mobile etc.), equally and efficiently. Whereas, in [40] CISCO clearly understands the marketing progress in IoT so it is working deliberately hard and proposed new technology named "FOG computing" as a solution to prominent problems. Its practical implementation is estimated up to 2018 by CISCO. In[41], TELIT as a service provider claims to be an IoT global enablement partner. Its focus is on Mobile Network Operation Management and security.

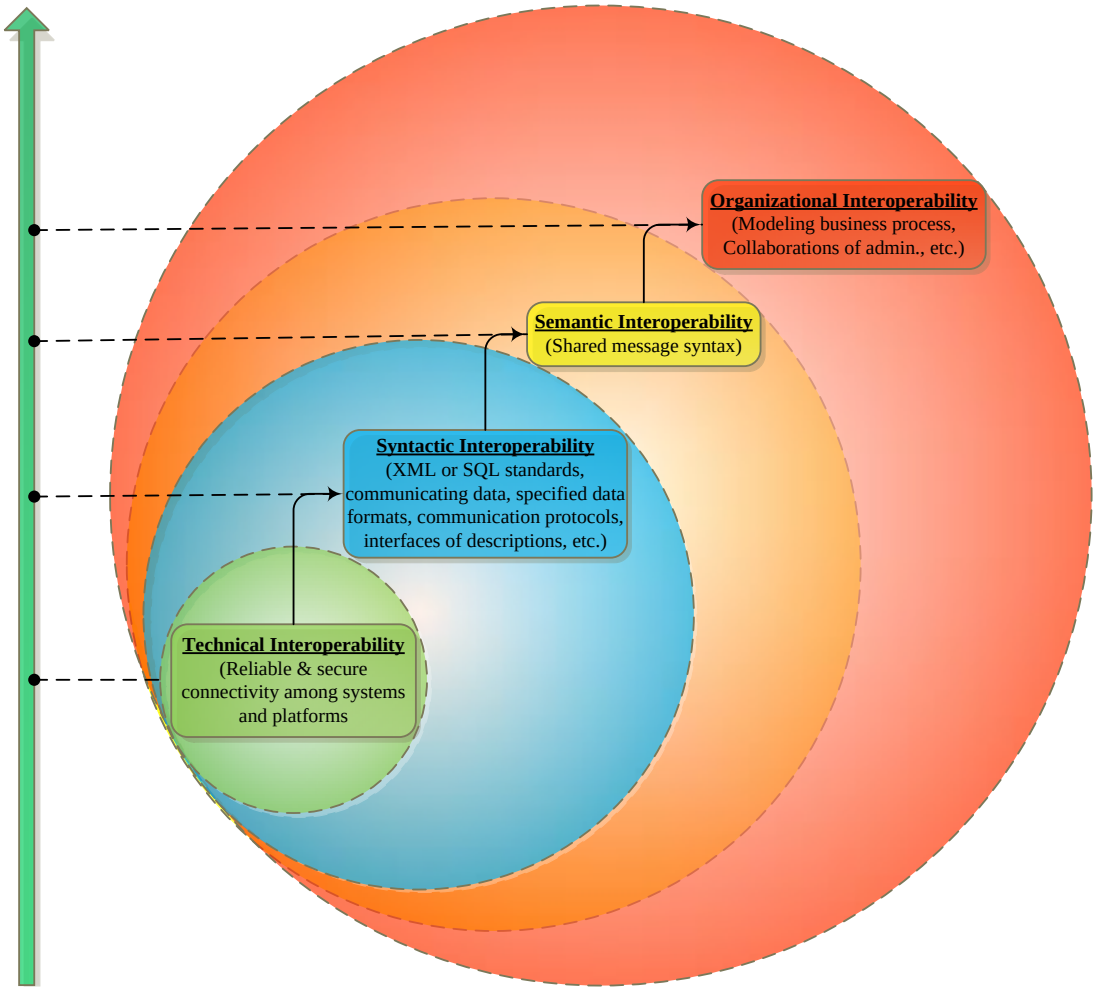


Fig. 4. Bottom-up Layered approach of Interoperability

H. Hardware Issue

Field of Nano-technologies has been quite evolved in recent years but still some software are very large to be handled on such level (in the IoT perspective) e.g., Linux with all features. Furthermore, the hardware issue with power and storage constraint also required to be managed. RFID technology has been researched a lot in this regard, which make system noticeable of small size and low cost [10][22][39][42]. It provides high radio coverage area. RFID sensor network can support computing data, communication and sensing devices abilities in an inactive system. In a real-time critical environment, IoT also needs back-end sensors, networks and infrastructure in case of any failure occurs in the regular IoT network. Now the IT business communities are more focusing on hardware development for the future market competition. CISCO [35], one of the competitive organizations; is claiming to produce such open hardware solution that will be able to work with other hardware components or devices. This means that these hardware solutions will be perfect solution for networking environments and it will provide sound functionality while connected to other devices. Moreover, there hardware and software solutions will be focused on security as security is the major issue in IoT. Such hardware solutions are needed for future Internet, which can handle all constraints mentioned above and new undiscovered problems because practical implementations are needed to test the IoT environments, which is currently not implementable due to certain limitations for example Smart City.

I. Interoperability

In [17], many organizations including ETSI (European Telecommunications Standards Institute), TIA (Telecommunications Indus- try Association), ITU-T (International Telecommunication Union), OMA (Open Mobile Alliance), GISFI (Global ICT Standardization Forum for India), CASAGRAS (coordination and support action for global RFID-related activities and standardization), CCSA(China communication standard association), etc. are specified. These organizations are specifically working for interoperability issue in IoT and M2M communication. M2M provide base for the IoT architecture as it describes the serviceable components of IoT. Interoperability is also addressed in [30][24][25][45][46] as one of the major key challenge. In [17], different levels of interoperability has been defined.

Figure 4 presents a bottom-up approach, which can also be described as layered approach. Each layer is dependent to the

layer below in some scenario, for example, syntactic interoperability is only possible, if technical interoperability exists so the syntactic interoperability will be the next step when technical interoperability is already implemented and so on. In [17], it is analyzed that there are two major issues regarding technical interoperability. One is the lack of complete reference architecture and other problem is the lack of technical interoperability estimation scope. Whereas, [47] claimed that technical interoperability can be achieved as technology progressing rapidly but business and semantic interoperability is the more challenging issues. Moreover, [48] stated that in technical area professionals needed to work with business area professionals as partners to achieve the IoT. Business interoperability is the most difficult challenge of all other interoperability.

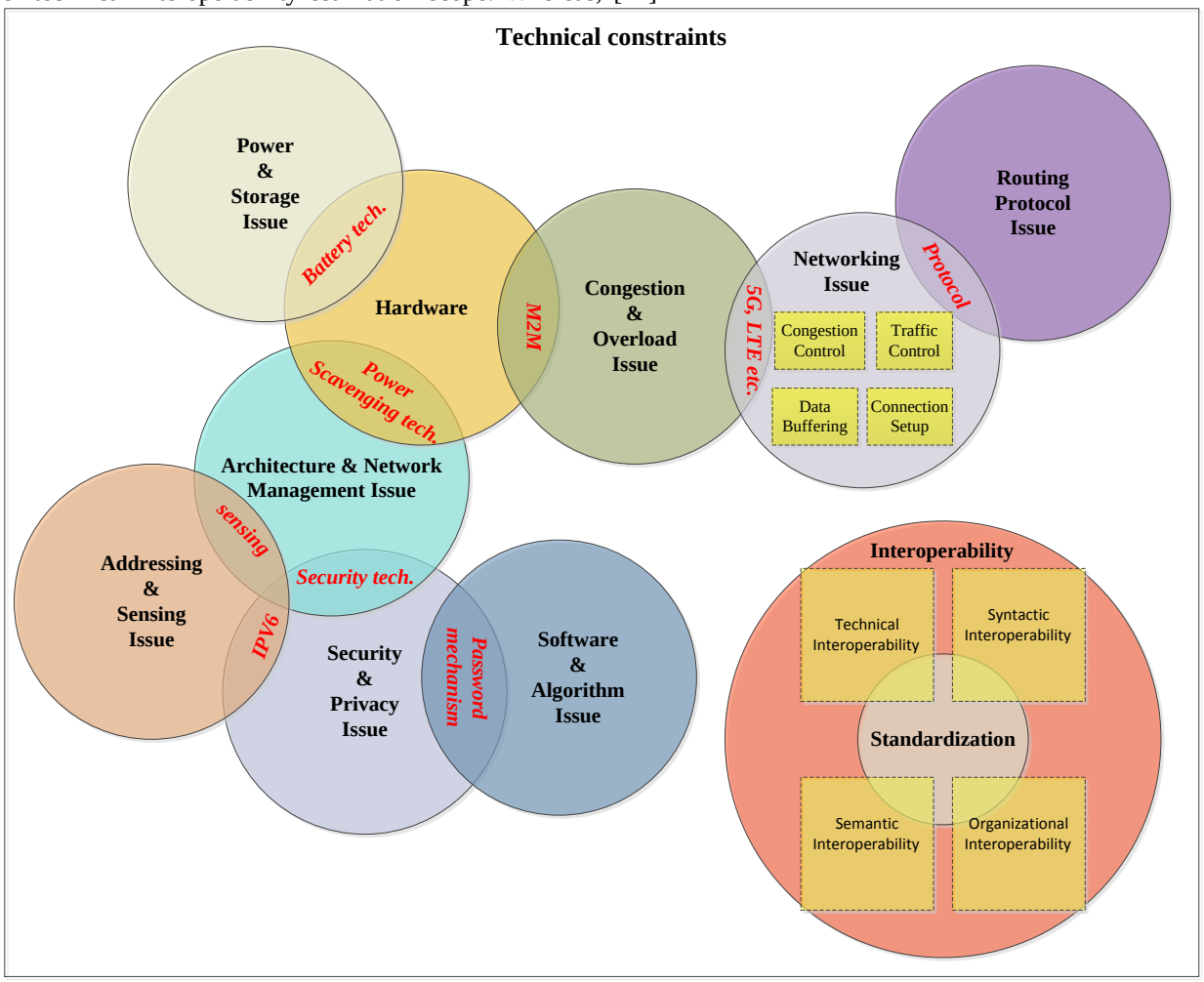


Fig. 5. Overlapping Technical Constraints

J. Congestion and Overload Issue

Congestion problem occurs when simultaneous messages came from multiple devices eventually leads to extreme overload situation which causes huge effect on network (3GP) which affects the network performance and leads to network failure. This situation can be seen in M2M and V2V communication and it has been researched by [5][6] related to the IoT. The congestion can also be occurred due to server or application malfunction. In [41], solution has been analyzed of congestion and overload issues, which can be resolved by LTE-advanced or existing technologies like LTE high bandwidth networks. One solution to this congestion control is to specify the time duration for connection. Devices can only connect to network when there is no overload and if the network is overloaded then disable all other connections. Second solution is to reject the connection from devices which create congestion problem.

K. Security and Privacy (Data transference issue)

Security and privacy is one of the most important hurdle of the IoT and it has been recently researched exhaustively by [49][9][50][20][51]. As security and privacy is a completely separate research area but indirectly collaborates with technical constraints therefore this survey does not focus on it but instead it just delivers the main idea. Technically, the IoT will not be applicable until or unless people tend to accept it. And this acceptance is correlated to the guarantee of their security and privacy. Data could not be collected in anticipation of the mistrust of people towards the Internet is cleared. However, the uncertainty of the security and privacy is due to the dark side of the IoT infrastructure. The future

internet will not only affect the IoT users but even non-users will also be targeted, indifferent to the todays Internet scenario.

The digital storage cost is tremendously decreasing, that result to store the information once generated to unlimited time, also included the fact about user's forgetting attitude towards digital data. Ultimately the IoT is providing an environment of great risk to privacy and security by integrating all of this digital data into world-wide network. Privacy and security can only be ensured if a user has fully control over his/her information. A user must know that what personal data is collected, who is collecting, where it is processed and when it was collected. Furthermore, the personal data should only be used by authorized service providers e.g., authorize medical organizations, authorize research institutions, authorize management systems, etc. Moreover, the data should only be stored for restricted time limit under severely need base scenarios otherwise it must be destroyed immediately. But this type of control is nearly impossible in sensor networks and the management authority is also difficult to define for this sort of control. Security constraint has also been focused by [6][18][26][52]. The data transference techniques are not able to control such level of security breaches due to technical limitations related to networking and middleware. Thus, there is a need to create such sort of data transference techniques that will not only handle high-level of security but also ensure authentication and data integrity.

The technical constraints explained in this section are overlapping each other but defined at some level of boundaries. Figure 5 highlighted this fact with the overlapping circles of constraints and red text in shaded areas in the overlapping regions. These overlapping regions show the interconnected behavior of the IoT technical constraints.

III. Most Recent Technological Developments

In this section, the most recent technological developments in the IoT perspective are briefly described. In [53] many recent technologies and their impact on daily lives have been analyzed. Customized services that matched for situation using location information, is implemented by using personal preferences and locating the person position. GPS and Google mapping is used to achieve the function. It can be used for different perspectives e.g., best nearby restaurants, hospitals, institutes, etc. With the use of smartphone, a person can be notified in nearby best restaurants, hospitals, institutions, service providers (Software houses), etc. according to personal preferences and customer or user received reviews.

The traditional Internet converts the world into a form of village but the future Internet is providing services entirely eliminating space constraints. It includes advancements such as smart city, precautionary maintenance system, remote electronics control service, etc. [53]. All electronics and devices are connected with social networking service (also called as social networking site or SNS). It manages real-time statistics e.g., temperature, speed, air pressure, and vibration. The statistics are then analyzed and prediction using data analytics are produced. In Smart City, the integrated CCTV control center recognized instead of human monitoring. If there has been a crime, the system guesses and investigates the suspect's expected escape routes. System notifies the police's mobiles with the related information. It minimizes the crime rate and prevents lives loss or additional damages. In precautionary maintenance system, engine failures in aircraft lead to flight delays due to repairs. This ultimately results in customer complaints at the end, and may even cause major accidents with casualties. For this reason, Rolls-Royce adopted a remote monitoring service by assigning sensors to their aircraft engines. In [46], Rolls-Royce also provides the preventive maintenance service, which eliminates probable intimidations based on its estimations. It could also lead the aircraft engine market by switching from a 'sales business': selling and supporting engines, to a 'service business': charging the rates on service used. Remote electronics control service provides Home Chat facility with the Electronics and mobile devices connected to home appliances. It enables a person to handle home appliances remotely.

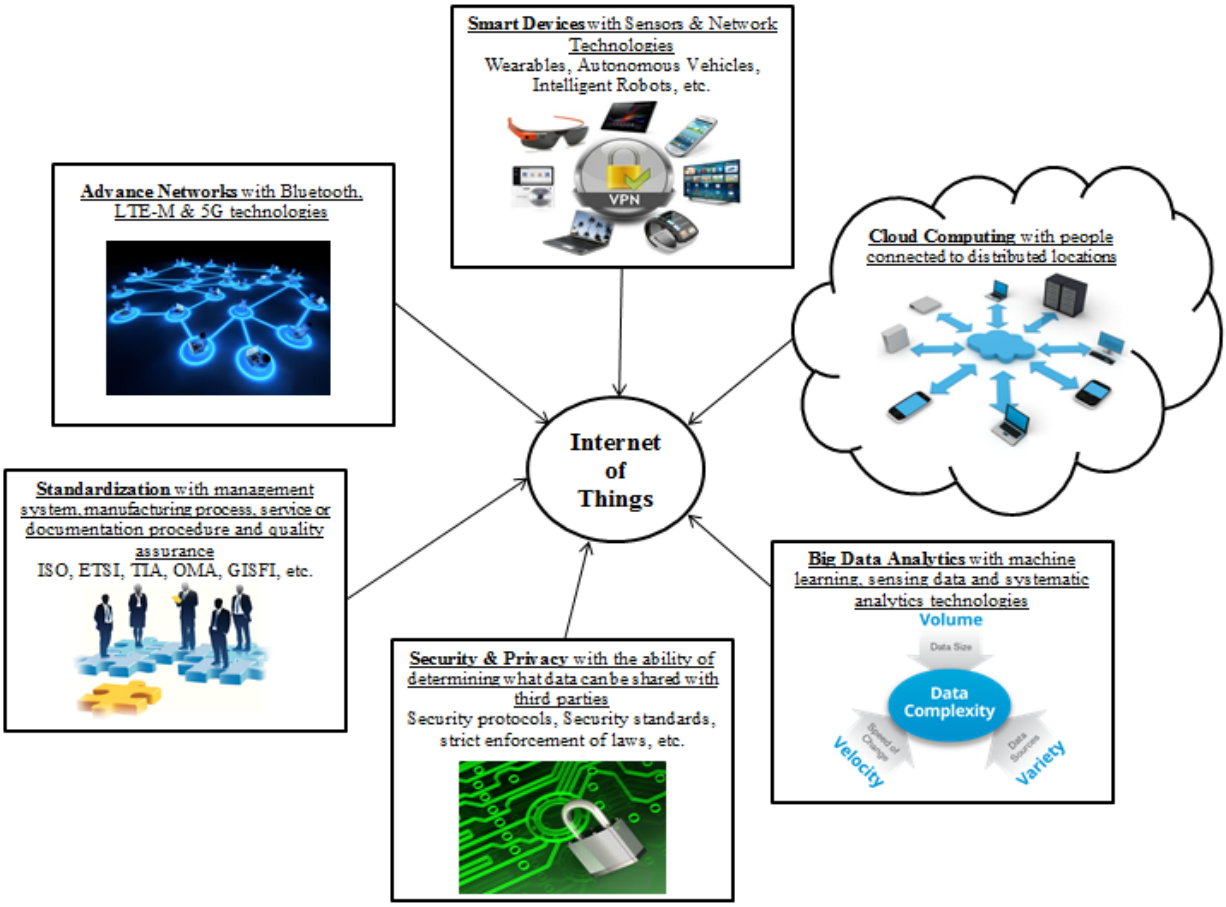


Fig. 6. Six key factors necessary to deploy the IoT

Optimized management systems with data-based computerization analyze data analytics. GE rail network optimization solution is an implemented system [54]. It reduces extra operational costs (such as a subway train slowing down to wait for the train, in front of it to be removed) because energy and time is needed to reduce speed sharply in keeping a safe distance from other trains and to recover speed. Railroad management is an implemented example of it. The aim of this system is that trains ride smoothly under a systematized system, which results in reducing extra operational costs. GE began monitoring all trains' operational status and locations to automatically determine the right speed for the schedule, with their rail network optimization solution and movement planner system. This change has improved their train speeds by 15 to 20% compared to the average [53].

Uber service is like the traditional call for taxis from the company. 'Transparency' is the key feature in the request process that results in privacy maintenance explained by [53]. Passengers have easy access to nearest available cab/taxi whenever and wherever they want. Personal information of the driver / passengers and reviews from previous passengers are maintained that ultimately provide better customer experience. Moreover, negative reviewers can be constrained from the Uber service for avoiding bad experiences. Smart Payment provides easy use of service.

LittleBits kit, is a hardware set for toys developed in the U.S. [55]. It is an electronic circuit development kit that enables people to develop the IoT lacking any low-level knowledge of development but only with little concept of Input and Output. This is one of the products and services that help even beginners to implement the IoT. The purpose of LittleBits is to provide a system where anyone can build, prototype, and learn about electronics. Each modular component is labeled with a purpose. By chaining them together with their magnetic links more complex circuits started to form. Hooking and unhooking are simple as magnets connect to each module. LittleBits kit can control all ranging from sound to real moving motors, powered by either a 9V battery or a power supply [56].

In Smart School [57], ongoing project involves varieties of multimedia equipment such as electronic pens, digital textbooks, tablet PCs, multimedia, etc. Using these high-tech instruments and integrating them into the system, with high-level connectivity desire functionality is achieved. Smart school based on basic key features of Smart Classes, Smart Information and Smart Management. In Smart class, an Integration of E-Learning environment is achieved by integrating High-Tech equipment. Smart card includes Multifunction (e-attendance, access control, and e-locker) and scalability as an e-cash service. Smart information provides various information such as school affairs, building locations

etc. and increasing security. Smart management remotely monitors and control devices/equipment and automated trouble warning. With such an integrated system, people in remote locations can share their judgments, heighten the quality of teaching resources, and make smart teamwork classrooms for a more proficient class.

TABLE I. TECHNOLOGICAL ADVANCEMENT IN THE IOT AND THE PERFORMANCE EVALUATION

Technology Advancement	Development Techniques	Category	Effect/Benefits on daily life	Limitation
Optimized management systems with data-based automation [53]	Integrated CCTV control centers Sensors and State controlling electronic devices Data analytics	Software Hardware Database management system	Organized system Reduce extra operational costs Improved train speeds by 15 to 20% compared to the average [53]. Network Optimization Solution and Movement Planner System.	Complex data analytics strategies, expensive
Customized services that suited for context using location information [53]	GPS Google mapping	Software Database Application	Notifications for nearby best service providers Personal preferences and customer or user received reviews.	End user info can be wrong
Creating new values by connecting objects [53]	GPS SNS Data analytics	Application	Transparency Improved customer service experience Smart payments and ease of access Avoid and control negativity to prevent bad events	Started to destroy the current market rapidly
Services eliminating space constraints [53]	Integrated CCTV control centers Sensors (temperature, air pressure, speed, vibration, etc.) Social networking service (SNS) Data analytics	Cyber-physical systems Software Management Application	Smart city: Replacement of human monitoring; System predictions expected escape routes; System fast notifications to police's mobile devices; Minimizes the crime rate and further costs. Preventive maintenance system: Minimize flight delays, customer complaints, and major accidents; Remote monitoring services; Preventive maintenance service and eliminates potential threats; Charged as per use. Remote electronics control service: Home Chat facility; Enables to handle home appliances remotely	Data management issue, can handle limited number of devices
E-Marketing / smart marketing [51]	GPS Credit cards	Data mining Database Management system	The return on investment (ROI) is quick; Customer experience improved; Easy Exchange of Sales Data; Instant Customer Analysis; Intelligent Devices That Know They're Dying; Analytical Social Media; Advertisements per interests/ preferences	Privacy issues, Lack of trust from both sides (customer & producer), Destroying current market
Smart School [57]	Varieties of multimedia equipment's High-tech instruments Wireless network	Software	Smart class: Integration of E-Learning environments Smart card: e-attendance, access control, e-locker and e-cash service Smart information: provides intelligent information and increased security Smart management: Remotely monitors and control and automated trouble warning. High quality sharing Enhance the quality of teaching materials	Social effect as due to emotions attachments: humans cannot be replaced by machines at some places
LittleBits [55]	Nano Tech. Sensors	Hardware	No low level hardware or software details are needed Facilitating beginner in development	Expensive
Oracle Solaris 11 [54] [55]	Virtual machines OS Physical domains High level connectivity	Virtualization Hardware Software Management system application	Fulfill business needs Increased efficiency High availability Elastic scalability Rapid deployment, development & management Economic (Pay as You Go) Low overhead & Dynamic Standard version and release of Oracle Solaris on all zones	Limited access control, Privacy & Security issue
Smart Manufacturing [66]	3D printing CAD or Scanning machines	Hardware Software	Cost effective manufacturing Customizable and automatic Time and energy saving Variety of materials are available	Intellectual property rights and criminal or illegal use

Autonomous vehicles [67]	Sensors Actuators High-level connectivity	Real time applications Hardware Database management system	Minimize human errors that can occur while driving Minimize death ratio or damages due to road accidents A person can drive without driving knowledge	Expensive, environmental changes effect performance and in case of accident: who to blame?
Driving Innovation in Health Systems through an Apps-Based Information Economy [68][69]	APIs Sensors Actuators GPS	Database management system Cyber-physical systems	APIs for cost effective health monitoring apps development (FHIR API, SMART API, Research Kit, Health Kit, Google Fit API, Validic API, 2net Platform, etc	Inter-operability
Emergency alert and communication system[10][71]	Sensors Actuators Data analytics	Mobile apps Augmented realties Hardware Management software	Improve relationship between public and private agencies Provide emergency reaction teaching to all employees, not just security workers Manage mechanism and authorizations with secure failover systems	Standardization and expensive
Smarter Highways variable speed limits [72][73]	Sensors actuators Data analytics	Real time application Cyber-physical devices	Automatic sensing Adjustable speed limits	Weather effects may affect the sensing
Smart Farming [9], [74]–[76]	Data analytics sensors	Real time application Cyber-physical devices	Increase productivity Reduce cost and economic Provide notifications before alarming situations	expensive
Wearable devices [77]	Data monitoring apps Sensors	Ubiquities computing Virtualization Hardware	Body stabilizers e.g., Kokoon in-ear sleep headphones High quality stimulated gaming Lifesaving monitoring Remote monitoring for pets Health alarming	expensive

The marketers experienced rapid increase in their sales by using E-Marketing [58]. It is beneficial for both service providers and customers. In E-marketing/smart marketing [59], High-level connectivity using GPS etc. are used. The return on investment (ROI) from E-Marketing can go beyond that of traditional marketing strategies and customer experience improved increasingly. The key features include easy exchange of sales data, Smarter CRM (customer relationship management) with instantaneous customer analysis, Intelligent Devices "*That Know They're Dying"* with their own regular maintenance and diagnostics, predictive social media for customer preferences or demands and 100% CTR (Click through Rate) with no phishing of advertisements. Instead, not only consumer will be saved from time wasting on irrelevant ads but service providers will also be facilitated by avoiding wasting of money on irrelevant ads.

Many other useful developments are in progress in a rapid speed. The top IT enterprises are in competition to achieve more market value in the IoT development race. In future, the IoT is more predictable to take over lives with its innovations.

IV. The IoT Necessities

In previous section, the recent technological advancements towards the IoT have been analyzed. Table I is providing brief analysis to these recent advancements while considering the related restrictions. By observing Table I, one can get an initial idea about the progress of the IoT and limitations. On the basis of Table I and II; it has been evaluated that the six key factors are necessary to deploy the Service-oriented Internet of Things. Four key factors to deploy the future Internet: Smart Devices, Advanced Networks, Cloud Computing and Big Data Analytics are already been described in [55]. However, the two most critical factors: Security and Standardization/Policies cannot be ignored. Therefore, this paper analyzes and refine the major key factors to the necessities of the IoT in Figure 6.

The technology advancement analyzed in these areas has improved energy consumption by 50% and increased the battery life by 50% [55]. That ultimately decreases the cost factor involved and rate of errors occurrence. The most important and huge impacts of the IoT on daily life has been analyzed by [53] is that the IoT is replacing jobs and it is switching the whole industrial structure. In Table 1, the technological progressions and their effects are analyzed. The advantages in the IoT world has several associated challenges whereas, the six basic principles for deploying the IoT has been highlighted in Figure 6

V. Performance Evaluation

The paper aimed to analyzing all the papers, books and articles for technical constraints from 2009 onwards. This section provides the tabular summary of recent research advancements. Table II presents the most stressed technical limitations of the IoT. Table II is made on the focus or requirements of the IoT and the technical constraints claimed by the recent researchers. Moreover, it provides a comprehensive summary of all the technical requirements and associated challenges currently being faced by the IoT. It will be helpful for all of the domain individuals related to directly or indirectly IoT.

TABLE II. COMPREHENSIVE OVERVIEW ABOUT TECHNICAL CHALLENGES OF THE IOT

IoT requirements	Technical challenges	IoT requirements	Technical challenges
Organizational Interoperability [45]	• Standards	Organizational Interoperability [44]	• Standards
E-governance [43]	• Semantic interoperability • Syntactic interoperability • Organizational interoperability • Technical Interoperability	[13]	• Need efficient and interoperable solutions • Cloud-based back end services • Adaptable and dynamic analytics solutions
[8]	• Network and security foundation • Size and scale of IoT providers	[6]	• Privacy, Identity Management, Security and Access control • Standardization and Interoperability • Data deluge
Smart devices [20]	• Scalability • "Arrive and operate" • Interoperability • Discovery • Software complexity • Data volumes • Data interpretation • Security and personal privacy • Fault tolerance • Power supply • Interaction and short-range communications • Wireless communications	[9]	• Identification Technology • Internet of Things Architecture Technology • Communication Technology • Network Technology • Software • Services and Algorithms • Hardware • Data and Signal Processing Technology • Discovery and Search Engine Tech. • Relationship Network Management Tech. • Power and Energy Storage Tech. • Security and Privacy Technologies • Standardization
Internet Protocol Wireless Sensor Network [19]	• IPV6 Adaptation • Mobility • Web-Enablement • Time synchronization • Security • No efficient communication protocol	IoT middleware [10]	• Open issue Standards • Mobility support Naming Transport protocol • Traffic characterization and QoS support • Authentication Data Integrity, Privacy, Digital forgetting
[49]	• Heterogeneity and Scalability • Security and Privacy • Search and Discovery • Ambient Intelligence	[21]	• Identification and Addressing • Internet scalability • Heterogeneity • Service Paradigms
IP smart objects and Service composition [78]	• Recursiveness • Semantic composition • Context-awareness • Hybrid composition • Privacy and security • Resource constraint • Power efficiency • Low-power & Lossy communication link • Data/event-driven services • Asynchrony • Discovery • Management requirements • QoS awareness	Cultural, ethical, socio-economic, but also technological expectations in-formation communication system [22]	• Processing and Handling Limitations • Storage Limitations • Transmission Limitations • Control Limitations • Traffic growth vs heterogeneity in capacity distribution • The current inter-domain routing system is reaching fundamental limits • Scaling to deal with flash crowding • Significant processing power / storage / bandwidth for indexing / crawling and (distributed) querying • Security of the whole Internet Architecture • Support of mobility
Smart home and smart building systems [26]	• Distributor-centric rather than customer-centric • Scalability	Circuits and Systems [25]	• Low power consumption • Highly modular approach • Diversity
[11]	• Standardization issue • Privacy and security issue • Routing protocol issue in V2V communication • Addressing and networking issue • Congestion and overload issue	Internal security and end-to-end user transparency [24]	• IPV6 transition
SmartCities, WSNs & M2M (constraint devices: low computation power, energy, memory)[7]	• Scalability • Governance • Lack of testbeds • Non-Interoperable solutions • No efficient new communication paradigm	[13]	• Need efficient and interoperable solutions • Cloud-based back end services • Adaptable and dynamic analytics solutions
Interoperability [47]	• Semantic interoperability	Light-weight IoT	• Semantic Interoperability

	• Syntactic interoperability • Organizational interoperability • Technical Interoperability	reference architecture [17]	• Syntactic Interoperability • Technical Interoperability • Organizational Interoperability
cloud computing [35]	• Security and cloud computing	M2M standards [28]	• Standardization • Interoperability
Application and usage [32]	• Privacy and security • Standardization	Internetworking [27]	• Congestion Control & Resource Allocation • Network Security
Smart environments [79]	• Standardization and policies	Device and data security [40]	• Nano-electronics • Devices secure management • Security algorithms
[80]	• Device and Data Security • centralized Service Management System(SMS) • secure remote management	RFID tech. [42]	• RFID technology transference to paper or plastic while holding the required productive resolution
[46]	• standardization and synchronization • Privacy • Pervasive and Trustworthy Network and Service Infrastructures • Nanotechnologies, sensor technologies, solutions bridging Nano and micro systems, etc. • Components, Systems, Engineering • Towards sustainable & personalized healthcare • Mobility, Environmental Sustainability & Energy Efficiency • Independent Living, Inclusion & Governance	Interoperability [29]	• Convergence in Technology • Integration of multiple data-sources • Unified Data Map / Ontology as point of reference • Mobility and Crowd sensing • P2P Communication • Data Modeling and Data Exchange • Ontology merging / Ontology matching & alignment • Data/Event Semantic Annotation • Knowledge Representation & related ontologies • Knowledge Sharing • Knowledge Revision & Consistency • Semantic Discovery of Data Sources, Data and Services • Semantic Publish/subscribe & Semantic Routing • Analysis & Reasoning
Plug n' Play smart objects [12]	• Complete Architecture • Efficient energy sensing • Secure reprogrammable networks and privacy • Quality of service (QoS) management • New protocols • Participatory sensing • Data mining for deep learning in terms of the need for adaptive, distributed and incremental learning techniques • GIS based visualization	Enabling technologies, Nano electronics, cyber physical systems, intelligent device management, smart gateways, telematics, smart network infrastructure, cloud computing, ecosystem and industrial applications [31]	• Security • Reliability • Complex integration • Discoverability • Interoperability • Standardization • High data rates • Dense crowds of users • Low latency • Low energy and • Low cost • A massive number of devices • Design of open APIs
Business view of IoT [51]	• Privacy, security and confidentiality • Standards • scalability	IoT Business and cloud computing [33]	• Standardization • Autonomic capabilities • Data operations • Privacy protection
[81]	• Heterogeneous and resource constrained devices • LLN(low power and lossy Network) • Confidentiality, mutual authentication & message origin authentication • Security protocols	Warehousing or future supply chain management [37]	• Integration • Agility • Consolidation • ROI • Standardization • performance guarantees and manager trust
Open systems for IoT [48]	• Data integration • Data automation • Data analyzation for identifying actionable insights • Scalability • Interoperability • Agility • Compatibility	Context-aware deployment of IoT [36]	• Automated configuration of sensors • Context discovery • Acquisition, modelling, reasoning, & distribution • Selection of sensors in sensing-as-a-service model • Security, privacy, and trust • Context Sharing
Security [52]	• Identification/ authentication • Trust • Reliability • Auto-immunity	[50]	• Bootstrapping • Mobility • Scalability • Data processing

	• Privacy • Responsibility • Safety		• Standardization • Protocol and network security • Data and privacy • Identity management • Trust and governance • Fault tolerance
[82]	• Security and privacy • Mobility management • QoS support • Protocols (at both network and transport layer) • Energy limitation and efficiency	[83]	• self-describable and self-contained • privacy and security • limited power supply • communication interoperability • semantic interoperability • syntactic interoperability
Business model [39]	• Devices Integration	IoT ecosystem [18]	• IoT business models and standards
Security [84]	• Routing protocol • Identity management framework		

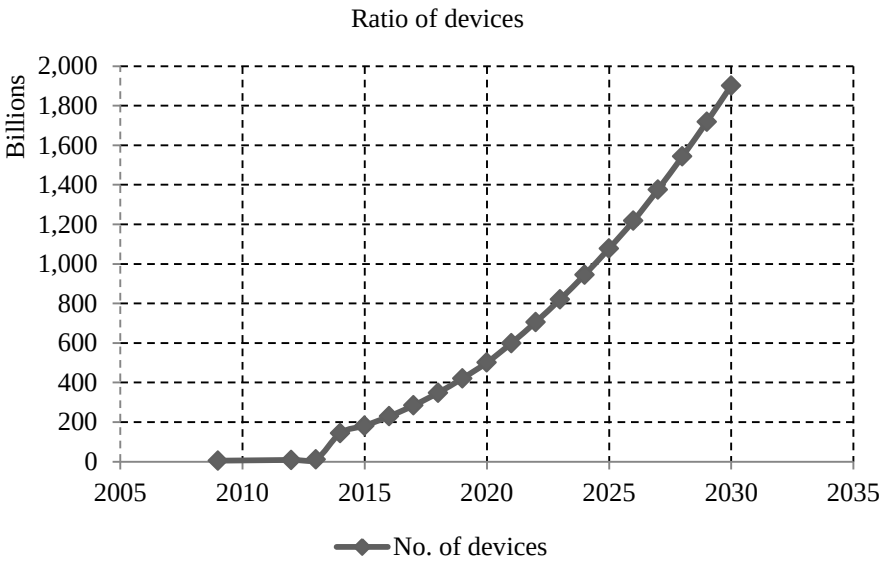


Fig. 7. Number of devices (in billions) per year

TABLE III. NUMBER OF DEVICES (IN BILLIONS) PER YEAR

Year	No. of devices in Billion
2009	4.8
2012	8.7
2013	11.2
2014	144
2015	182
2016	229
2017	284
2018	348
2019	421
2020	501
2021	599
2022	706
2023	821
2024	945
2025	1,078
2026	1,218
2027	1,376
2028	1,543
2029	1,718
2030	1,902

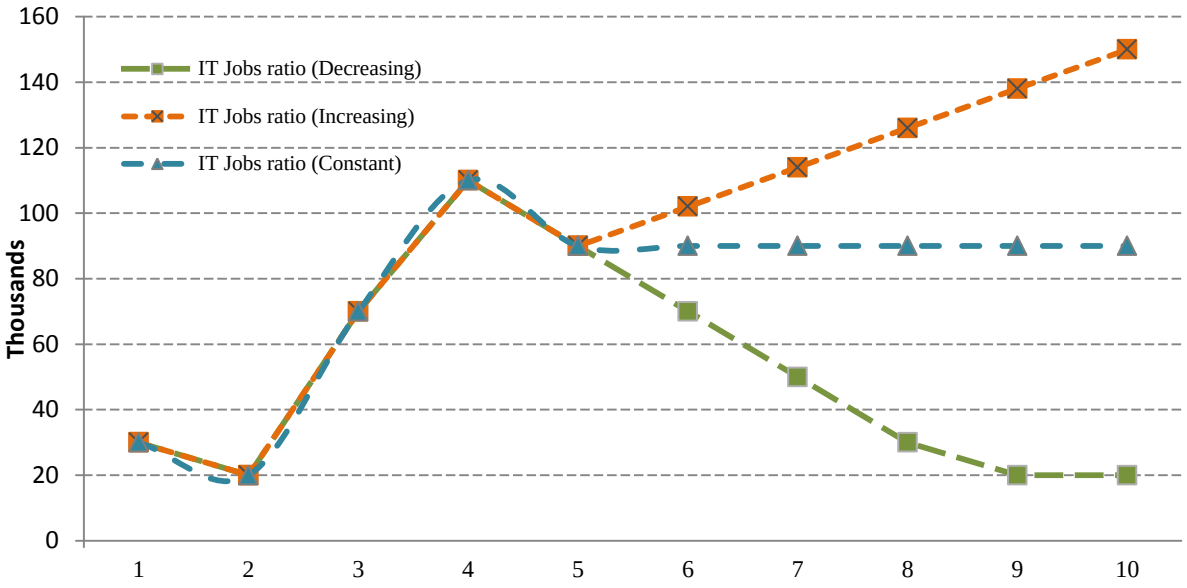


Fig. 8. Estimated number of IT jobs per year from 2011 to 2020

VI. The World in 2020 and Beyond

By analyzing details in previous sections, it can be concluded that the IoT is important enough to change the way of living and even the whole world. The technical constraints analyzed in this paper are the most important and critical ones. These technical constraints are key factors for the future Internet and can enormously affect cost factor for the future devices and technology. The affect the social life style and the way of living in many perspectives will be greatly influenced. An IT research company, Gartner; estimated the human resource demand will be reduced up to 50% by 2018 due to the intelligent machines technology advancements. With the high-speed development in the IoT and its job formation effect, Digital Business related jobs are also predicted to grow up 50%. Repetitive and life-critical work can be replaced by the IoT services. Moreover, Automation functionalities as a key feature of the IoT made it possible to take over data analysis. The industrial structure is creating new values by changing the whole infrastructure and destroying the traditional system revenues. Enterprise value of Uber was analyzed as $45 billion in December of 2014. The Wall Street Journal has analyzed the abrupt growth and estimated it to be $50 billion in May of 2015. The opportunity of existing industrial dominators could be threatened by novel outstanding ideas' innovators, is getting higher.

TABLE IV. Number of Jobs (in Number) Per Year

Year	IT Jobs ratio (Decreasing)	IT Jobs ratio (Constant)	IT Jobs ratio (Increasing)
2011	30000	30000	30000
2012	20000	20000	20000
2013	70000	70000	70000
2014	110000	110000	110000
2015	90000	90000	90000
2016	70000	90000	102000
2017	50000	90000	114000
2018	30000	90000	126000
2019	20000	90000	138000
2020	20000	90000	150000

If the ratio of devices connectivity is closely observed up to year: 2020 then the number of devices in next year's up to 2026 can be estimated. Figure 7 shows the prediction on the common increasing ratio among previous years' estimations. However, the Table III shows the statistics for assumptions of this paper; the gray shading color highlights the calculated assumptions. Indirectly, the number of devices per year is affecting the number of IT jobs per year. In contrast, the IT jobs are not completely down falling but their average value is not increasing as well. Figure 8 demonstrates the estimated ratio of IT jobs over years. It can be increasing, decreasing or constant amount of IT jobs according to the observation of

previous years' ratio. Overall analysis for the future Internet era clearly delivers the fact that successful companies or organizations will be those, who will provide some innovative technological ideas. The others will be just doomed out economically.

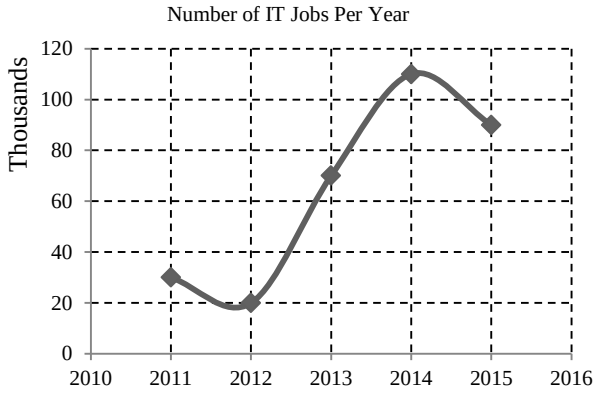


Fig. 9. Number of jobs (in number) per year

The statistical data obtained from [60] shows the number of IT jobs in recent years (up to 2015). The data has been analyzed to estimate the ratio of IT jobs in future (up to 2020). The aforesaid ratio has been plotted in Figure 9 whereas, Table IV shows the statistics. The three variations of IT jobs in Figure 9 shows the three assumptions for increasing (highlighted by green color), decreasing (highlighted by orange color) or constant rate (highlighted by green color) of IT jobs. These assumptions are being calculated on average ratios of recent years. The quick analysis of IT jobs against advancement in IT shows that average rate of IT jobs is down falling and not increasing with the rate of the IoT advancements. The reason behind down falling of IT jobs are vague but can be justified by keenly analyzing technology advancements. With the rapid growth in automation and artificial intelligence properties in industries eventually minimizing the human resource factor needed to operate and maintain the environments. The industrialization became more successful entity as compare to other government level jobs. That results in privatization effect, which ultimately goals more work with fewer jobs.

Another major factor involved for discussed downfall is the increased ratio of population. Moreover, young and educated people belong from urban areas or less developed areas moves to rural or developed areas. This ultimately increases the competition for better opportunities as per the life standard are changed. The IoT as advancement in technology is also playing major role in changing life standards of society.

VII. Conclusion

The concept of the IoT is maturing rapidly and soon we will be seeing the world-wide interconnected network, integrating every physical device with other devices. The IoT has been focused exhaustively in recent years due to its wider applications such as eliminating space, reducing cost, saving energy and intensive monitoring. This paper highlighted the core areas of the IoT and specifically targeted the technical constraints that act as critical hurdle in the way to deploy the successful IoT infrastructure. We categorized these challenges in security, hardware, standards, gateway systems, middleware and the IoT databases etc. Open issues to the IoT are also enlightened that are required to be addressed by the researchers and other stakeholders of the IoT. Moreover, our future predictions will help the concerned parties to prepare for the IoT accordingly. In future, we aim to provide solutions about the open issues to the IoT.

References

[1] S. N. Han, I. Khan, G. Myoung, N. Crespi, and R. H. Glitho, "Computer Standards & Interfaces Service composition for IP smart object using realtime Web protocols : Concept and research challenges," Comput. Stand. Interfaces, vol. 43, pp. 79–90, 2016.

[2] CompTIA, "Sizing Up the Internet of Things," 2015.

[3] Swan, "Sensor Mania ! The Internet of Things , Wearable Computing , Objective Metrics , and the Quantified Self 2 . 0," J. Sens. Actuator Networks, pp. 217–253, 2012.

[4] S. Taylor, "How service providers can help businesses to realize the promise of the IoT revolution," CISCO, 2016. [Online]. Available: http://blogs.cisco.com/sp/how-service-providers-can-help-businesses-to-realize-the-promise-of-the-iot-revolution#more-185149. [Accessed: 09-Oct-2016].

[5] Bauer, M. Boussard, N. Bui, F. (UniS) Carrez, C. (SIEMENS) Jardak, J. (ALUBE) De Loof, C. (SAP) Magerkurth, S. Meissner, A. (FhG I. Nettsträter, A. (CEA) Olivereau, M. (SAP) Thoma, W. W. Joachim, J. (CSD/SUni) Stefa, and A. (UniWue) Salinas, "Internet of Things – Architecture IoT-A Deliverable D1 . 5 – Final architectural reference model for the IoT v3 . 0," 2013.

[6] L. Coetzee and J. Eksteen, "The Internet of Things – Promise for the Future ? An Introduction," IST-Africa Conf. Proc., pp. 1–9, 2011.

[7] Elmangoush, H. Coskun, S. Wahle, and T. Magedanz, "Design aspects for a reference M2M communication platform for Smart Cities," in 2013 9th International Conference on Innovations in Information Technology (IIT), 2013, pp. 204–209.

[8] Lopez Research, "An Introduction to the Internet of Things (IoT)," Lopez Res. Llc, vol. Part 1. of, no. November, pp. 1–6, 2013.

[9] H. Sundmaeker, P. Guillemin, and P. Friess, Vision and challenges for realising the Internet of Things, no. March. 2010.

[10] L. Atzori, A. Iera, and G. Morabito, "The Internet of Things: A survey," Comput. Networks, vol. 54, no. 15, pp. 2787–2805, 2010.

[11] S. Agrawal and D. Vieira, "A Survey on Internet of Things : Security and Privacy Issues," Abakós, vol. 1, pp. 78–95, 2013.

[12] J. Gubbi, R. Buyya, S. Marusic, and M. Palaniswami, "Internet of Things (IoT): A vision, architectural elements, and future directions," Futur. Gener. Comput. Syst., vol. 29, no. 7, pp. 1645–1660, 2013.

[13] P. Barnaghi and A. Sheth, "Internet of Things: The Story So Far," 2014.

[14] Brandon Butler, "12 most powerful Internet of Things companies | Network World," 2014. [Online]. Available: http://www.networkworld.com/article/2287045/wireless/153629-10-most-powerful-Internet-of-Things-companies.html. [Accessed: 17-Dec-2015].

[15] Skerrett, "How to Categorize the Internet of Things - DZone IoT." [Online]. Available: https://dzone.com/articles/how-categorize-internet-things. [Accessed: 16-Dec-2015].

[16] Gigli, "Internet of Things: Services and Applications Categorization," Adv. Internet Things, vol. 1, no. 2, pp. 27–31, 2011.

[17] S. Bandyopadhyay, P. Balamuralidhar, and A. Pal, "Interoperation among IoT Standards," J. ICT Stand., vol. 1, no. 2, pp. 253–270, 2013.

[18] S. Leminen, M. Westerlund, M. Rajahonka, and R. Siuruainen, "Towards IOT ecosystems and business models," Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics), vol. 7469 LNCS, pp. 15–26, 2012.

[19] G. Deak, K. Curran, J. Condell, E. Asimakopoulou, and N. Bessis, "IoTs (Internet of Things) and DfPL (Device-free Passive Localisation) in a

disaster management scenario," Simul. Model. Pract. Theory, vol. 35, no. October 2015, pp. 86–96, 2013.

[20] F. Mattern and C. Floerkemeier, "From the internet of computers to the internet of things," Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics), vol. 6462 LNCS, pp. 242–259, 2010.

[21] S. Haller, S. Karnouskos, and C. Schroth, "The Internet of things in an enterprise context," Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics), vol. 5468, pp. 14–28, 2009.

[22] H. Schaffers, A. Sallstrom, M. Pallot, J. M. Hernandez-Munoz, R. Santoro, B. Trousse, J. Domingue, A. Galis, A. Gavras, T. Zahariadis, D. Lambert, F. Cleary, P. Daras, S. Krco, H. Müller, M.-S. Li, and others, The Future Internet-Future Internet Assembly 2011: Achievements and Technological Promises. 2011.

[23] Stockebrand, "IPv6 Address Basics," in IPv6 in Practice: A Unixer's Guide to the Next Generation Internet, Berlin, Heidelberg: Springer Berlin Heidelberg, 2007, pp. 21–34.

[24] G. Han, C. Bao, X. Li, and S. Liu, "IPv6 Transition for the Other Billions," Comput. Commun. Networks (ICCCN), 2015 24th Int. Conf., 2015.

[25] Blaauw, D. Sylvester, P. Dutta, Y. Lee, I. Lee, S. Bang, Y. Kim, G. Kim, P. Pannuto, Y.-S. Kuo, D. Yoon, W. Jung, Z. Foo, Y.-P. Chen, S. Oh, S. Jeong, and M. Choi, "IoT design space challenges: Circuits and systems," in 2014 Symposium on VLSI Technology (VLSI-Technology): Digest of Technical Papers, 2014, pp. 1–2.

[26] Spano, L. Niccolini, S. Di Pascoli, and G. Iannacconeluca, "Last-Meter Smart Grid Embedded in an Internet-of-Things Platform," IEEE Trans. Smart Grid, vol. 6, no. 1, pp. 468–476, 2015.

[27] B. S. Peterson, Larry L and Davie, Computer Networks: A Systems Approach, 5th ed., 5th ed. Morgan Kaufmann, 2011.

[28] T. Klinpratum, C. Saivichit, A. Elmangoush, and T. Magedanz, "Performance of Interworking Proxy for Interconnecting IEEE1888 Standard at ETSI M2M Platforms," Appl. Mech. Mater., vol. 781, pp. 141–144, 2015.

[29] W. Pollard, IoT Semantic Interoperability: Research Challenges, Best Practices, Recommendations and Next Steps. IERC (European Research Cluster On The Internet Of Things), 2015.

[30] Microsoft Corporation, "Ten reasons your business needs a strategy to capitalize on the Internet of Things today," 2014.

[31] L. and others Medagliani, P and Leguay, J and Duda, Andrzej and Rousseau, Franck and Duquennoy, S and Raza, S and Ferrari, Gianluigi and Gonizzi, P and Cirani, S and Veltri, Internet of Things Applications - From Research and Innovation to Market Deployment. River Publishers, 2014.

[32] Chaouchi and Hakima, The Internet of Things: Connecting Objects. John Wiley & Sons, 2013.

[33] S. Shen and M. Carug, "An EvolutionaryWay to Standardize the Internet of Things," J. ICT Stand., vol. 2, no. 2, pp. 87–108, 2014.

[34] H. Petersen, M. Lenders, M. Wählisch, O. Hahm, and E. Baccelli, "Old Wine in New Skins? Revisiting the Software Architecture for IP Network Stacks on Constrained IoT Devices," p. 6, Feb. 2015.

[35] J. Vanian, "Cisco CEO Chuck Robbins talks importance of the Internet of things - Fortune," 2015. [Online]. Available: http://fortune.com/2015/10/05/cisco-chuck-robbins-internet/. [Accessed: 15-Dec-2015].

[36] C. Perera, A. Zaslavsky, P. Christen, and D. Georgakopoulos, "Context Aware Computing for The Internet of Things: A Survey," vol. X, no. X, pp. 1–41, 2013.

[37] J. Reaidy, A. Gunasekaran, and A. Spalanzani, "Bottom-up approach based on Internet of Things for order fulfillment in a collaborative warehousing environment," Int. J. Prod. Econ., vol. 159, pp. 29–40, 2015.

[38] N. Meghanathan, S. Boumerdassi, N. Chaki, and D. Nagamalai, Eds., Recent Trends in Networ b k Security and Applications, vol. 89. Berlin, Heidelberg: Springer Berlin Heidelberg, 2010.

[39] TIMMS, "Monitoring 4.0: How to track the Internet of Things | Business Spectator," 2014. [Online]. Available: http://www.businessspectator.com.au/article/2014/6/11/technology/monitoring-40-how-track-internet-things. [Accessed: 02-Dec-2015].

[40] S. David Lake, Ammar Rayes, and Monique Morrow, "The Internet of Things - The Internet Protocol Journal, Volume 15, No. 3," 2013. [Online]. Available: http://www.cisco.com/web/about/ac123/ac147/archived_issues/ipj_15-3/153_internet.html. [Accessed: 15-Dec-2015].

[41] "Subscriptions & Rate Plans - Telit," 2015. [Online]. Available: http://www.telit.com/products-and-services/iot-connectivity/subscriptions-rate-plans/. [Accessed: 16-Dec-2015].

[42] Scd. E. Staff, "RFID and AIDC News: New Chipless RFID Tag Could Transform the Industry," 2015. [Online]. Available: http://www.scdigest.com/ontarget/15-04-22-1.php?cid=9228. [Accessed: 12-Dec-2015].

[43] H. Kubicek, R. Cimander, and H. J. Scholl, Organizational Interoperability in E-Government. Springer Science \& Business Media, 2011.

[44] H. Gottschalk, Petter and Solli-Saether, "Levels of Organizational Interoperability," E-Government Interoperability Inf. Resour. Integr. Fram. Aligned Dev. IGI Glob. Philadelphia, PA, pp. 242–244, 2009.

[45] H. Kubicek and R. Cimander, "Three dimensions of organizational interoperability. [Insights from recent studies for improving interoperability frame-works," Eur. J. ePractice, no. January, pp. 1–12, 2009.

[46] Toma, E. Simperl, and G. Hench, A joint roadmap for semantic technologies and the internet of things, vol. 1, no. APRIL. 2009.

[47] Gottschalk and H. Solli-Sæther, E-Government Interoperability and Information Resource Integration. IGI Global, 2009.

[48] N. Noronha, Andy and Moriarty, R and Connell, KO and Villa, "Attaining IoT Value : How To Move from Connecting Things to Capturing Insights Gain an Edge by Taking Analytics to the Edge," Cisco Anal. Br., 2014.

[49] Zeng, S. Guo, and Z. Cheng, "The Web of Things: A Survey (Invited Paper)," J. Commun., vol. 6, no. 6, 2011.

[50] Roman, P. Najera, and J. Lopez, "Securing the Internet of things," Computer (Long. Beach. Calif)., vol. 44, no. 9, pp. 51–58, 2011.

[51] Uckelmann, M. Harrison, and F. Michahelles, "Architecting the Internet of Things," pp. 1–25, 2011.

[52] Riahi, Y. Challal, E. Natalizio, Z. Chtourou, and A. Bouabdallah, "A Systemic Approach for IoT Security," 2013 IEEE Int. Conf. Distrib. Comput. Sens. Syst., pp. 351–355, 2013.

[53] "What Can 'Things' do When Connected to 'The Internet'? -Talk Service-Oriented IoT(2)- | LG CNS Blog | Creative & Smart." [Online]. Available: http://www.lgcnsblog.com/features/what-can-things-do-when-connected-to-the-internet-talk-service-oriented-iot2/. [Accessed: 16-Dec-2015].

[54] "Train keeping promises, smart rail system | GE reports Korea." [Online]. Available: http://www.gereports.kr/ge-trip-optimizer-for-smart-train-solution/. [Accessed: 17-Dec-2015].

[55] "What We Need to Implement IoT - Talk Service-Oriented IoT (3) - | LG CNS Blog | Creative & Smart." [Online]. Available: http://www.lgcnsblog.com/features/what-we-need-to-implement-iot-talk-service-oriented-iot-3/. [Accessed: 16-Dec-2015].

[56] "How to Get Started DIYing Anything with LittleBits." [Online]. Available: http://lifehacker.com/how-to-get-started-diying-anything-with-littlebits-1617311793. [Accessed: 21-Dec-2015].

[57] "A Step Closer to Educational Equality with ICT and Smart Schools | LG CNS Blog | Creative & Smart." [Online]. Available: http://www.lgcnsblog.com/features/a-step-closer-to-educational-equality-with-ict-and-smart-schools/. [Accessed: 16-Dec-2015].

[58] "Advantages and disadvantages of online marketing." [Online]. Available: http://zeendo.com/info/advantages-and-disadvantages-of-online-marketing/. [Accessed: 21-Dec-2015].

[59] Leung, "5 Ways the Internet of Things Will Make Marketing Smarter - Salesforce Blog," 2014. [Online]. Available: https://www.salesforce.com/blog/2014/03/internet-of-things-marketing-impact.html. [Accessed: 17-Dec-2015].

[60] "Number of IT Jobs down." [Online]. Available: http://www.e-janco.com/Press/2015/2015-09-04-Forecast-New-IT-Jobs-Revised.html. [Accessed: 17-Dec-2015].

[61] Health and I. Portability, "SANS Institute InfoSec Reading Room," 2001.

[62] M. Rouse, "What is OSI reference model (Open Systems Interconnection)? - Definition from WhatIs.com," 2014. [Online]. Available: http://searchnetworking.techtarget.com/definition/OSI. [Accessed: 09-Oct-2016].

[63] C. Borysowich, "EA Deliverable: Architecture Strategy: Interoperability," 2008. [Online]. Available: http://it.toolbox.com/blogs/enterprise-solutions/ea-deliverable-architecture-strategy-interoperability-sample-27850. [Accessed: 09-Oct-2016].

[64] "Building a Cloud-Based Data Center with Oracle Solaris 11 - Part 1." [Online]. Available: http://www.oracle.com/technetwork/articles/servers-storage-admin/build-cloud-solaris11-2172575.html. [Accessed: 21-Dec-2015].

[65] Enterprise, O. Distribution, and N. Virtualization, "Oracle Solaris 11 - Engineered for Cloud | Oracle."

[66] "Smart Manufacturing magazine to Launch in Spring 2016 -- DEARBORN, Mich., Dec. 7, 2015 /PRNewswire-USNewswire/ --." [Online]. Available: http://www.prnewswire.com/news-releases/smart-manufacturing-magazine-to-launch-in-spring-2016-300188473.html. [Accessed: 25-Jan-2016].

[67] "CES 2016: Carmakers kick off the year with big moves in autonomous vehicles - TechRepublic." [Online]. Available: http://www.techrepublic.com/article/ces-2016-carmakers-kick-off-the-year-with-big-moves-in-autonomous-vehicles/. [Accessed: 25-Jan-2016].

[68] K. D. Mandl, J. C. Mandel, and I. S. Kohane, "Driving Innovation in Health Systems through an Apps-Based Information Economy," Cell Syst., vol. 1, no. 1, pp. 8–13, Jun. 2015.

[69] "Internet of Things Examples - Postscapes." [Online]. Available: http://postscapes.com/internet-of-things-examples/. [Accessed: 25-Jan-2016].

[70] "Overcoming Emergency Notification Challenges in 2016." [Online]. Available: http://www.everbridge.com/overcoming-emergency-notification-challenges-in-2016/. [Accessed: 25-Jan-2016].

[71] "AtHoc - Networked Crisis Communication." [Online]. Available: http://www.athoc.com/. [Accessed: 25-Jan-2016].

[72] "Smarter Highways variable speed limits." [Online]. Available: http://www.wsdot.wa.gov/smarterhighways/vsl.htm. [Accessed: 25-Jan-2016].

[73] R. No and P. For, "Variable speed limit signs effects on speed and speed var i ation in work," no. January, 2008.

[74] "Reading Beehives: Smart Sensor Technology Monitors Bee Health and Global Pollination | Libelium." [Online]. Available: http://www.libelium.com/temperature-humidity-and-gases-monitoring-in-beehives/. [Accessed: 25-Jan-2016].

[75] "Smart Farming: Monitoring Horses and Equine Facility Management with Waspmote | Libelium." [Online]. Available: http://www.libelium.com/smart-farming-monitoring-horses-equine-facility-management-waspmote. [Accessed: 25-Jan-2016].

[76] "Sustainable Farming and the IoT: Cocoa Research Station in Indonesia | Libelium." [Online]. Available: http://www.libelium.com/sustainable-farming-and-the-iot-cocoa-research-station-in-indonesia. [Accessed: 25-Jan-2016].

[77] "50 wearable tech gamechangers for 2016." [Online]. Available: http://www.wareable.com/wareable50/best-wearable-tech. [Accessed: 25-Jan-2016].

[78] N. Han, I. Khan, G. M. Lee, N. Crespi, and R. H. Glitho, "Service Composition for IP Smart Object using Realtime Web Protocols: Concept and Research Challenges," Comput. Stand. Interfaces, vol. 43, pp. 79–90, Aug. 2015.

[79] William Pollard, Internet of Things: Pan European Research and Innovation Vision. IERC, 2011.

[80] Strom, "The Demise of Web 2.0 and Why You Should Care," Internet Protoc. J., vol. 15, no. 3, pp. 20–24, 2012.

[81] J. Park, S. Shin, and N. Kang, "Mutual Authentication and Key Agreement Scheme between Lightweight Devices in Internet of Things," vol. 38, no. 9, 2013.

[82] Iera, C. Floerkemeier, J. Mitsugi, and G. Morabito, "Guest Editorial: The internet of things," IEEE Wirel. Commun., no. December, pp. 8–9, 2010.

[83] M. Jazayeri, S. Liang, and C.-Y. Huang, "Implementation and Evaluation of Four Interoperable Open Standards for the Internet of Things," Sensors, vol. 15, no. 9, pp. 24343–24373, Sep. 2015.

[84] D. Meghanathan, Natarajan and Boumerdassi, Selma and Chaki, Nabendu and Nagamalai, Recent Trends in Network Security and Applications. Springer, 2010.

ETEEM- Extended Traffic Aware Energy Efficient MAC Scheme for WSNs

Younas Khan[1]
Department of Computer Sciences
Institute of Management Sciences
Peshawar, Pakistan

Sheeraz Ahmed[3]
Department of Electrical Engg
University of Engg & Technology
Peshawar, Kohat, Pakistan

Saqib Shahid Rahim[5]
Department of Computer Sciences
Abasyn University
Peshawar, Pakistan

Fakhri Alam Khan[2]
Department of Computer Science
Institute of Management Sciences
Peshawar, Pakistan

Imran Ahmad[4]
Department of Computer Sciences
Institute of Management Sciences
Peshawar, Pakistan

M. Irfan Khattak[6]
Department of Electrical Engg
University of Engg & Technology
Peshawar, Kohat, Pakistan

***Abstract*—Idle listening issue arises when a sensor node listens to medium despite the absence of data which results in consumption of energy. ETEEM is a variant of Traffic Aware Energy Efficient MAC protocol (TEEM) which focuses on energy optimization due to reduced idle listening time and much lesser overhead on energy sources. It uses a novel scheme for using idle listening time of sensor nodes. The nodes are only active for small amount of time and most of the time, will be in sleep mode when no data is available. ETEEM reduces energy at byte level and uses a smaller byte packet called FLAG replacing longer byte SYNC packets of S-MAC and SYNCrts of TEEM respectively. It also uses a single acknowledgement packet per data set hence reducing energy while reducing frequency of the acknowledgment frames sent. The performance of ETEEM is 70% better comparative to other under-consideration MAC protocols.**

Keywords—Energy Consumption; Multi-hop; Network Allocator Vector; Throughput; Wireless Sensor Networks

I. INTRODUCTION

Wireless sensor networks (WSNs) [1] are used in scenarios where human body cannot reach easily like earthquakes, battlefield, flood forecasting or for weather forecasting etc. WSN is a wireless network consisting of autonomous devices or nodes using sensors to monitor physical or environmental changes [2]. Each wireless node consists of sensor [3], radio transceiver [4], microcontroller [5], buffer [6] and power source [7]. Sensor is used to sense data in proximity. Transceiver is used to receive and send data to other sensor nodes. When data is sensed by sensor, then node becomes active (i.e. its state changes from sleep state to active state), processes that data and forwards it as specified by protocol. WSN uses adhoc topology [8] and most of the time the nodes are scattered on geographical area. Each node uses its own power source to operate. Wireless sensor nodes may use multi hop communication [9], in which one sensor node may pass data to another which in turn may forward it to other nodes till it reach to base station. Power consumption of wireless nodes depends on node operations. Power is critical to communication [10]. Four major sources of power wastage [10] in WSN are idle listening time, Overhearing, Overhead and Hidden terminal problem.

Idle listening problem arises when a sensor node listens to medium despite the absence of data which results in consumption of energy. Overhearing occurs in multi hop communication in which a node senses data which is destined to another node. Similarly, overhead occurs as a result of overhearing.

Several techniques, such as S-MAC [10], T-MAC [11], RMAC [12] TEEM [13] etc have been used to optimize energy consumption of WSN node. Sensor-MAC (S-MAC) was proposed to minimize energy consumption for all four sources (i.e. idle listening time, overhearing, overhead, hidden terminal problem) of power wastage. One major drawback of S-MAC protocol is its fixed time interval of active and sleep state which may still result in idle listening time. To address this issue, Traffic Aware Energy Efficient MAC (TEEM) protocol was proposed. TEEM protocol uses variable time for active and sleep state depending upon traffic instead of fixed interval. TEEM protocol reduces energy from idle listening and overhead part of energy consumption sources. When there is data with the nodes then it is active for whole life cycle else it goes to sleep state very earlier then its normal schedule. Both S-MAC and TEEM protocols send Acknowledgment (ACK) packet per fragment. TEEM consumes 16% less energy than S-MAC; but still it wastes energy in SYNCdata part i.e. if there is no data then it still waits for SYNCnodata part. Further, ACK packet sent per fragment also consumes energy. Due to these issues in TEEM protocol and S-MAC protocols we developed Extended Traffic Aware Energy Efficient MAC (ETEEM) protocol. ETEEM protocol not only incorporates variable time slot for active and sleep interval, but also introduces ACK packet per burst instead of ACK packet per fragment; resulting in reduced overhead. It uses four- byte packet called FLAG instead of ten byte SYNC packet. Major contributions of this paper include:

1) *Energy optimization due to reduced idle listening time*
2) *Energy optimization due to reduced overhead*
3) *Implementation of ETEEM protocol, simulation and evaluation*

4) Comparison of ETEEM protocol with S-MAC protocol and TEEM protocol

II. LITRATURE REVIEW

Initially 802.11 wireless standards IEEE [14] were developed to deal with wireless networks. CSMA/CA [15] under 802.11 standards was considered as the standard protocol. CSMA/CA is used for mobility (cellular) in wireless networks. It uses RTS/CTS and ACK technique for contention for the medium. It cannot be used in WSN because this standard gives no service for power saving and scalability for multi hop communication.

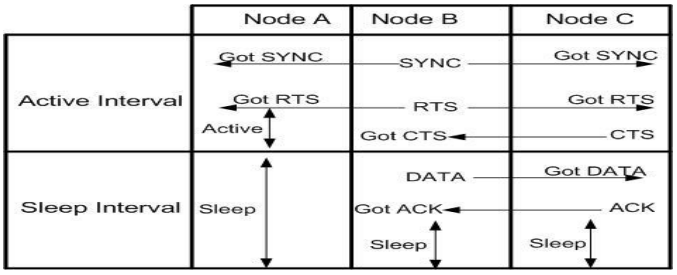


Fig. 1. S-MAC-When node has some data to send

In S-MAC protocol, sensor node first sends SYNC packet to form virtual clustering and CTS/RTS/ACK technique for contention for the medium, and subsequently makes communication with the right user. If a sender occupies the medium, next step is to send RTS packet, RTS packet has the corresponding receiver node address with which sender wants to communicate. When RTS reaches to receiver, it finds its address inside the packet and recognizes that sender node wants to communicate with him. As a result receiver node sends CTS packet. CTS mean that receiver node is ready to catch sender packets. Figure 1[13] shows S-MAC protocol behavior when there is data with one node and it want to send it to other node. Nodes will remain active in sleep state if data arrives and will go to sleep state as soon as data transfer is completed. It behaves differently when node does not have any data transfer with other nodes as showed in figure 2 [13].

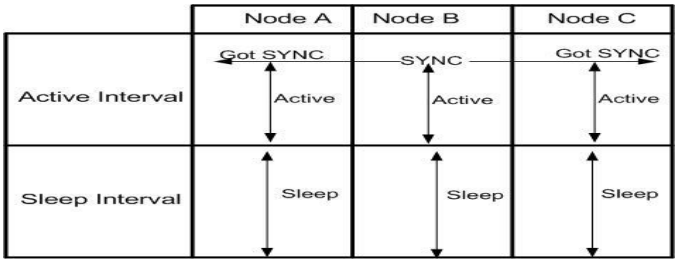


Fig. 2. S-MAC-When node has no data to send

S-MAC divides the whole life of each node into two parts: Listen and Sleep interval. In listen interval, wireless nodes can take part in communication with other nodes; and in sleep interval, it turns off its transceiver and does not listen to any data unless sleep time expires. Listen interval of sensor node is further divided into two parts: SYNC and DATA as in figure 2. If a node joins the network, it waits for random amount of time. However, if it does not listen to any schedule broadcasted by other nodes it can choose its own schedule and broadcast it in SYNC packet to its immediate neighbors, which further broadcasts it and eventually reaches to all nodes.

S-MAC uses fixed time intervals for listen and Sleep state. When node is in active state, it sends data in burst but if a node does not has any data to send then it must wait for whole listen interval to go to sleep state. Due to fixed time interval, it consumes energy because nodes wait for the whole listen time to over. T-MAC protocol, inspired by S-MAC protocol takes work further and makes communication between two nodes in WSN depends on time. T-MAC protocol does not uses fixed active and sleep interval. It allows nodes to stay awake until an activation event has occurred for certain amount of time. Apart from this, in T-MAC protocol nodes sends data in burst and go to sleep state between bursts which minimize idle listening time. It also uses FRTS technique. FRTS stands for Future Request To Send. T-MAC protocol increases overhead by including FRTS mechanism which consumes energy. If two or more nodes send FRTS frame at same time then there is possibility of collision which waste energy.

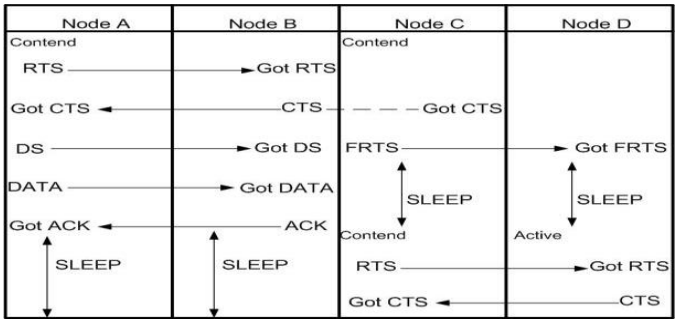


Fig. 3. FRTS frame of T-MAC

R-MAC protocol inspired by FRTS technique uses a special packet called POIN instead of RTS and CTS packet. Like S-MAC it divides nodes life cycle in three parts: SYNC, DATA and SLEEP. R-MAC protocol ensures that nodes will send data in sleep period. POIN is initiated by multi-hop which informs all nodes on path as when to wake up in sleep period. R-MAC protocol improves latency as compared to S-MAC and T-MAC protocol but it cannot guarantee collision avoidance. Two nodes when initiate POINS may succeed in transmitting POINS but both of them will start communication to their intended receiver at the start of sleep period and hence collision may occur. It wastes large percentage of energy in collision. TEEM [13] makes periodic awake and sleep intervals depending on traffic. The basic function is when there is traffic with node, then it sends it in burst in SYNCdata , and if there is no data to sends then it only send SYNC packet and goes to sleep state. Hence, it can minimize half the power compared to S-MAC. TEEM protocol divides sensor node lifecycle into two parts: SYNCdata and SYNCnodata. Nodes can send data in SYNCdata part and if does not have any data to send, then goes to sleep state in SYNCnodata part; too early then Sleep state as in figures 4 and 5. It also uses SYNCrts packet in contention for medium. RTS packet combined with SYNC is used to minimize overhead.

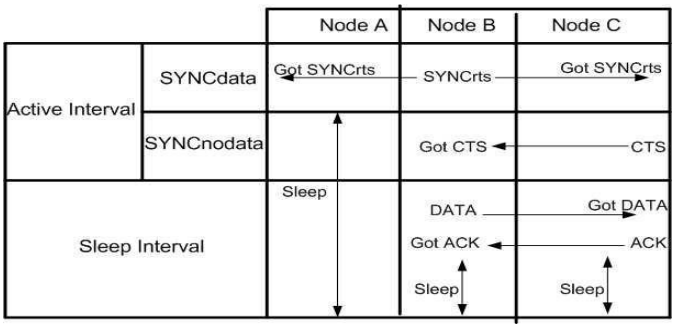


Fig. 4. TEEM-When node has some data to send

From previous figure, it is clear that TEEM protocol saves energy from idle listening time and overhead sources of energy consumption. But if we look closely, TEEM protocol also wastes energy in idle listening. If nodes do not have any data to send, then they waste first half that is SYNCdata part in idle listening. It is so because nodes will trigger their sleep schedule in SYNCnodata part.

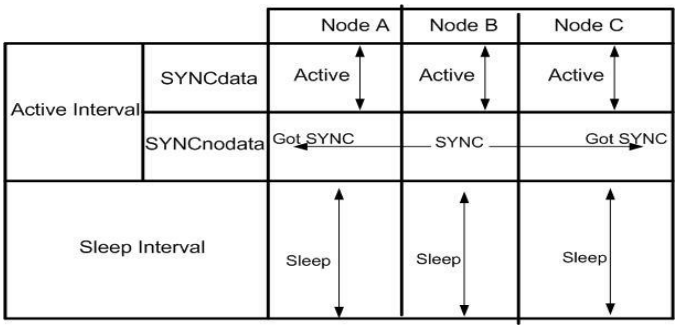


Fig. 5. TEEM-When node has no data to send

III. Proposed ETEEM Model

The proposed framework ensures that wastage of energy consumed by previous protocols in various ways is minimized. It saves energy from idle listening and overhead part of power consumption sources. ETEEM protocol neither has SYNCnodata, SYNCdata parts like TEEM protocol nor fixed listen and sleep time like S-MAC protocol. When node sleeping quantum expires and has no data to transfer or receive, then it waits for random amount of time and sends FLAG with value 0. If it has data to send, then it sends SYNCrts packet. In ETEEM protocol, nodes will not wait any more and will not listen to idle medium. S-MAC protocol sends separate SYNC and RTS packets which consume too much energy. It also sends ACKs per fragment which waste energy. TEEM protocol combines SYNC and RTS packet called as SYNCrts packet as shown in figure 5 [15]. Through this technique TEEM minimizes wastage of energy but TEEM protocol also supports ACKs per fragment technique which waste too much energy like S-MAC. ETEEM protocol, on one hand sends SYNCrts packet combining SYNC and RTS packet hence saves energy. On the other hand, it sends only one ACK for whole burst. There is no need to send ACKs per fragment. ETEEM protocol saves energy from this part also. The whole burst will be acknowledged in one ACKs packet.

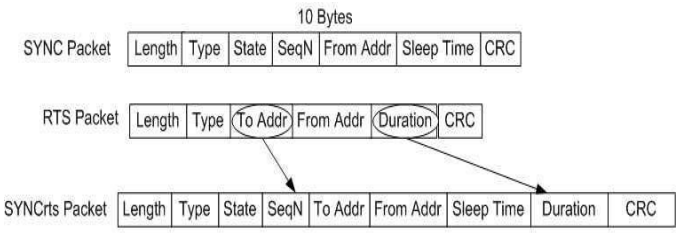


Fig. 6. SYNCrts packet

1) SYNCrts

WSN uses adhoc topology some nodes may die or disconnected while some may join later. Therefore, every wireless sensor network protocol should be capable to handle Synchronization. ETEEM protocol support Synchronization the same way as TEEM does i.e. through SYNC packet. In ETEEM protocol, when new node joins topology, it waits for some time and listens to the medium to see any of its neighbors broadcasted its schedule or not. If it hears schedule broadcasted by its neighbor, it simply adopts its schedule and stores it in schedule table. If there is no schedule broadcasted by its neighbor then it chooses its own listen and sleeps time and broadcast it to its immediate neighbors.

2) Adding Extra Packet

TEEM protocol divides Listen interval of nodes in two parts: SYNCdata and SYNCnodata. If node has data to send then it sends it in SYNCdata part, and if it does not have any data to send then it waits for SYNCnodata part. In SYNCnodata part of listen interval it sends SYNC packet to its immediate neighbor indicating that it has no data and go to sleep state. If we examine closely, TEEM protocol wastes first half i.e SYNCdata part. I f node does not have data to send, then it just consumes energy while sending no data to other nodes. ETEEM enhances this portion of TEEM protocol. ETEEM uses one extra packet, called FLAG having two fields: Data and Length as shows in Figure 6. Data has two possible values either 0 or 1. If Data has 0 value, it means node has no data, if it is 1 then the receiver must check the length field which will show the duration of the traffic.

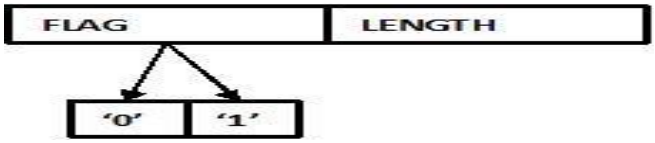


Fig. 7. FLAG packet with values 0 and 1

If node sleep time expires, it waits for random amount of time. After certain time, if it does not have any data to send, it will send FLAG packet with data value 0 and go to sleep state. If it has data to send, it will send RTS packet. After receiving CTS packet from the receiver, normal communication begins. After last packet, if it has further data arrived from other nodes and wants to send to the same receiver, then it will send FLAG with data value 1 indicating that it has more data to send. Receiver will resend RTS packet and further communication starts again. The receiver resends RTS packet for its neighbor nodes. It is possible that its neighbors sleep time expires and come to listen state, then through RTS packet

they will come to know that medium is still busy. Therefore, they will again go to sleep state, otherwise, will send FLAG with Data value 0 indicating no further data and will go to sleep state early than its schedule, hence saving energy. It consumes less energy as compared to S-MAC and TEEM as depicted in figures 8 to 10.

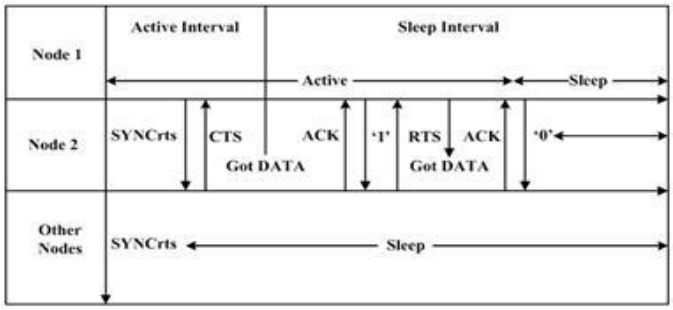


Fig. 8. when there is further data to send

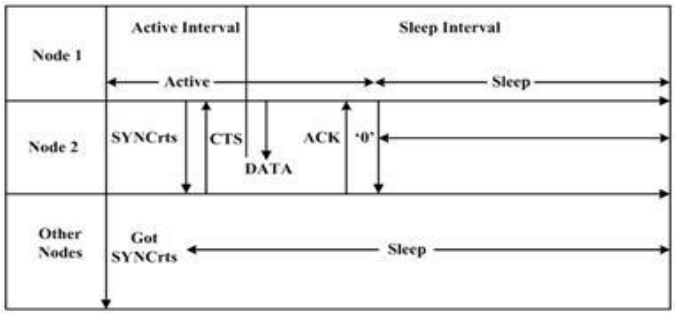


Fig. 9. when there is data with nodes

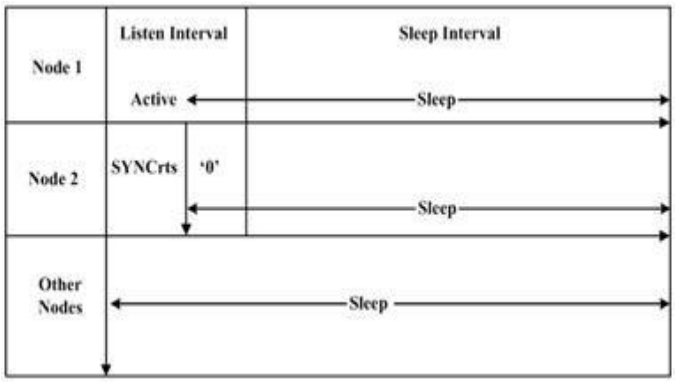


Fig. 10. When no data with nodes

3) Acknowledgment

Both S-MAC and TEEM protocols send ACK packet after each fragment received by receiver. They do so to avoid collision at receiver due to hidden terminal problem. Through ACK packet both protocols ensure receiver neighbors that medium is still busy. But it wastes energy too, due to overhead. ETEEM protocol sends one ACK packet for the whole burst i.e. for all fragments. It minimizes energy consumption because of sending ACK packet per whole burst. After missing packet received by the receiver, it acknowledges whole burst. There is possibility of collision at receiver. ETEEM compromises one collision per communication but has very little chances to occur as experimental result shows. If it occurs, it still consumes less energy from sending ACKs per fragment. On one collision, less energy is consumed and only one packet will be lost which can be recovered later. Collision can be minimized as in S-MAC and TEEM protocols but penalty will be paid in the form of energy consumption.

4) Algorithm for ETEEM

ETEEM will follow the work flow and algorithm as shown in figure 11. The figure shows that if sender wants to send data to receiver it first sends SYN Crts packet after receiving CTS both parties establishes connection. After last packet received with FLAG 0 it sends missing fragment sequence number to the sender. Sender will send again missing packets. If last packet is received and FLAG is 1 the receiver will first send lost packet sequence number and then sends CTS packet again. The algorithm will start working again from point 4.

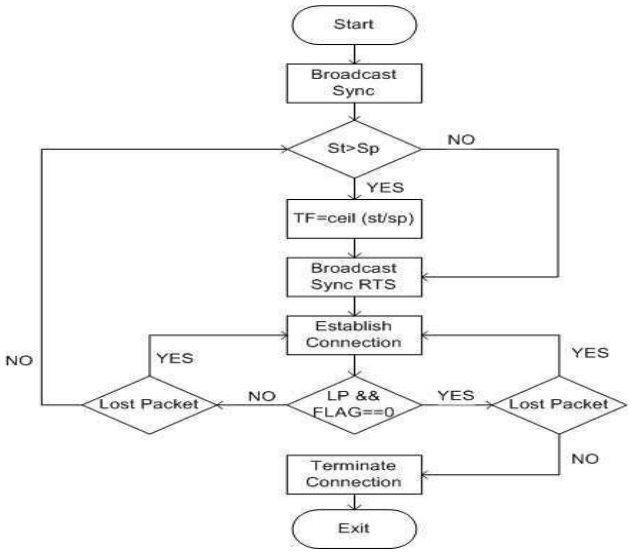


Fig. 11. Workflow of ETEEM protocol

IV. Performance Analysis

We took 03 nodes and simulation time as 50 seconds but it can be changed according to needs. Further, for SYNC packet and control packets, we took real values of energy consumption from the simulation of S-MAC and TEEM. Assumed that each node can send and receive 10 bursts of data and can change as per need. We implemented S-MAC, TEEM and ETEEM for three possible conditions i.e best case, average case and Worst case and calculated results. In order to evaluate these cases we make data as constant for all these protocols.

We divided our experiment in two parts: in first part, we devised formula for these protocols based on their algorithms and working was implemented for all possible conditions i.e. for Best Case, Average Case and Worst case. Experiment shows that ETEEM consumes less energy as compared to S-MAC and TEEM. S-MAC was the first protocol to implement in WSN.

1) S-MAC Protocol

We first took single node and calculated algorithm of S-MAC and then implemented that calculation on N number of

nodes. For topology creation and further contention for the medium Node first sends SYNC packet and Control Packets, therefore we have as in equation (1) below

$$N = SYNC + ControlPacket \dots\dots\dots\dots\dots\dots\dots\dots\dots\dots (1)$$

where ControlPacket = RTS + CTS hence equation (1) becomes

$$N = SYNC + RTS + CTS \dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots (2)$$

After successful transmission of control packets, data transmission will occur, so which is generating the same results as of equation (5)

$$N = SYNC + RTS + CTS + Data \dots\dots\dots\dots\dots\dots\dots\dots\dots (3)$$

As we know that Data = ACK + Fragment because after each fragment S-MAC also acknowledges. Hence equation (3) becomes

$$N = SYNC + RTS + CTS + ACK + Fragement \dots\dots (4)$$

As Data is variable, one node can send as many data fragment as it has in its buffer and after each fragment it also receives an ACK. After putting this scenario in equation (4)

$$N = SYNC + RTS + CTS + \sum_{x=1}^{n} (ACK + Data) \dots\dots\dots (5)$$

Similarly for total number of nodes N_t the operation will be described as follows

$$N_t = \sum_{y=1}^{n} (SYNC + RTS + CTS) + \sum_{x=1}^{N} (ACK + Data) \dots (6)$$

2) TEEM Protocol

TEEM protocol is modified version of S-MA C i.e. SYNCrts in place of SYNC. In order to devise formula for TEEM protocol we will have to first look to single node and then will implement it on N number of Nodes. For single node

$$N = SYNCrts + CTS + \sum_{x=1}^{N} (ACK + Data) \dots\dots\dots\dots\dots (7)$$

For n number of nodes the formula will be like following

$$N_t = \sum_{y=1}^{n} (SYNCrts + CTS) + \sum_{x=1}^{N} (ACK + Data) \dots\dots\dots (8)$$

3) ETEEM Protocol

As ETEEM is modified form of TEEM protocol, hence it uses FLAG as extra packet and uses one ACK for whole burst of data. For topology generation it also uses SYNCrts packet. For single node algorithm of TEEM protocol:

$$N = SYNCrts + CTS + \sum_{x=1}^{n} (ACK + Data + FLAG) \dots (9)$$

For total number of Nodes Nt, it will be:

$$N_t = \sum_{y=1}^{n} (SYNCrts + CTS) + \sum_{x=1}^{n} (ACK + Data + EOD) (10)$$

V. Performance Cases

1) Case 1

In Case 1, ETEEM protocol is compared with S-MAC and TEEM protocol for the scenario in which all these protocols consume their energy up to their high level as its algorithm support. We make a congested environment and force these protocols to collide each and every fragment sent by any node to its neighbor. We observed that nodes are active for their whole life cycle. For each fragment loss, S-MAC and TEEM protocols send ACKs twice i.e. one negative and one positive ACKs. ETEEM protocol sends only two ACKs i.e. one negative ACKs and one positive per burst. Experiment are depicted in figure 12 which shows that ETEEM consumes 0.0035% and 0.0043% less energy from TEEM and S-MAC respectively

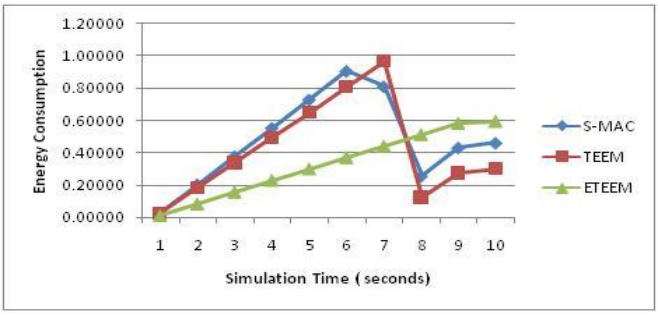


Fig. 12. Case 1 Scenario

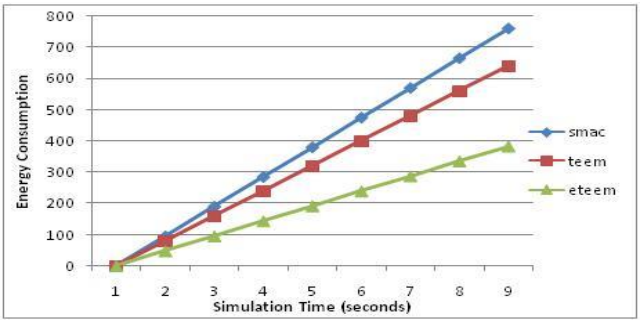


Fig. 13. Case 2 Scenario

2) Case 2

Case 2 is the scenario in which there is average traffic going on. We allowed large amount of data to be transferred in such a way that data transfer can occur in sleep quantum. All protocols are active for specified time equally depending on the data. But as S-MAC uses SYNC and RTS packets separately therefore it consumes more energy. TEEM combined these two packets and save some part of the energy but like S-MAC it acknowledges each and every data fragment. In collision S-MAC and TEEM almost send ACKs twice per collide fragment as compared to their normal routine. The performance of ETEEM is observed better than these protocols because despite the fact that it is active for the time as TEEM is, yet it sends only two ACKs i.e. one negative ACK and one positive, hence saves more energy than S-MAC and TEEM. Figure 13 shows that in average case ETEEM consumes 0.0032% and 37.5% less energy as compared to TEEM protocol and S-MAC respectively.

3) Case 3

Case 3 is the scenario in which it consumes minimal energy. As S-MAC uses fixed interval of time for Listen and Sleep schedule. Therefore Case 3 of S-MAC will be to go to sleep state within given schedule. Similarly for TEEM will be to go to sleep state in SYNCnodata part of Listen interval i.e. go early to sleep state. For ETEEM it will be the one in which it send FLAG packet with value 0 and go to sleep state early then its normal schedule. Best case result of these protocols are depicted below in figure 14 which shows that in best case ETEEM consume 70% less energy than TEEM protocol and subsequently consumes 92.5% less energy.

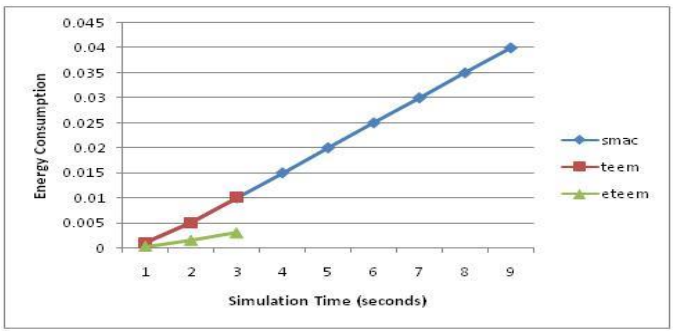


Fig. 14. Case 3 Scenario

VI. Conclusions and Future Work

This paper presents new MAC protocol for WSNs called ETEEM. It has very good energy conserving mechanism as compared to S-MAC and TEEM. Experimental result shows that it consumes less energy in all three possible cases (best case, average case and worst case). ETEEM protocol uses FLAG packet instead of SYNCnodata or SYNCdata part of TEEM. It uses one ACK per whole burst for the sake of energy conservation. For experimental purpose C language is used. Energy comparison is showed for three possible cases i.e. Best case, Average case and worst case. ETEEM consumes less energy in all three cases discussed in previous section. Future work includes experimentations of this protocol in different simulations tools. Nodes and Data can be varied to examine optimum results.

References

[1] Akyildiz, Ian F., et al.”Wireless sensor networks: a sur vey.” COMPUT NETW 38.4, 2002: 393-422.

[2] Chong, Chee-Yee, and Srikanta P. Kumar. ”Sensor networks: evolution, opportunities, and challenges.” Proceeding s of the IEEE 91.8, 2003: 1247-1256.

[3] Akyildiz, Ian F., Dario Pompili, and Tommaso Melodia.” Underwater acoustic sensor networks: research challenges.” Ad hoc networks 3.3, 2005: 257-279.

[4] Polastre, Joseph, Robert Szewczyk, and David Culler. ” Telos: enabling ultra-low power wireless research.” LECT NO TES COMPUT SC, Information Processing in Sensor Networks, 2005. IPSN 2005. Fourth International Symposium on. IEEE, 2005.

[5] Javaid, N., Ahmad, A., Rahim, A., Khan, Z. A., Ishfaq, M., and Qasim, U. (2014). Adaptive medium access control protocol for wireless body area networks. INT J DISTRIB SENS N, 2014.

[6] Alkhatib, Ahmad Abed Alhameed, and Gurvinder Singh Baicher. ”Wireless sensor network architecture.” International conference on computer networks and communication systems (CNCS 2012) Vol. 35. 2012.

[7] Philipp Sommer,Roger Wattenhofer Lars Schor, ”Towards a zero-configuration wireless sensor network architecture for smart buildings,” in First ACM Workshop on Embedded Sensing Systems for Energy-Efficiency in Buildings, 2009, pp. 31-36.

[8] Lu, Chenyang, et al. ”Rap: A real-time communication arc hitecture for large-scale wireless sensor networks.” Real -Time and Embedded Technology and Applications Symposium, 2002. Proceedings. Eighth IEEE. 2002.

[9] Anastasi, Giuseppe, et al. ”Energy conservation in wire less sensor networks: A survey.” Ad Hoc Networks 7.3, 2009: 537-568.

[10] Ye, Wei, John Heidemann, and Deborah Estrin. ”An energy -efficient MAC protocol for wireless sensor networks.” IEEE INFOCOM SER 2002. Twenty-First Annual Joint Conference of the IEEE Computer and Communications Societies. Proceedings. Vol. 3. 2002.

[11] Koen Langendoen Tijs van Dam, ”An adaptive energy efficient MAC protocol for wireless sensor networks,” ACM, pp. 171 - 180, 2003.

[12] Du, Shu, Amit Kumar Saha, and David B. Johnson.” RMAC: A r outing-enhanced duty-cycle MAC protocol for wireless sensor networks.” IEEE INFOCOM SER 2007. 26th IEEE International Conference on Computer Communications. IEEE. IEEE, 2007.

[13] Changsu Suh and Young-Bae Ko, ”A traffic aware energy effi cient MAC protocol,” IEEE, vol. 3, pp. 2975-2978, May 2005.

[14] Crow, Brian P., et al. ”IEEE 802.11 wireless local area networks.” Communications Magazine, IEEE 35.9, 1997: 116-1 26.

[15] Eusthatia Ziouva, ”CSMA/CA performance under high tra ffic condition: throughput and delay analysis,” computer communication, vol. 25, pp. 313-321, May 2002.

Intelligent System for Detection of Abnormalities in Human Cancerous Cells and Tissues

Jamil Ahmed Chandio, M. Abdul Rahman Soomrani
Department of Computer Science
Sukkur Institute of Business Administration (SIBA)
Sukkur, Pakistan

***Abstract*—Due to the latest advances in the field of MML (Medical Machine Learning) a significant change has been witnessed and traditional diagnostic procedures have been converted into DSS (Decision Support Systems). Specially, classification problem of cancer discovery using DICOM (Digital Communication in Medicine) would assume to be one of the most important problems. For example differentiation between the cancerous behaviours of chromatin deviations and nucleus related changes in a finite set of nuclei may support the cytologist during the cancer diagnostic process. In-order to assist the doctors during the cancer diagnosis, this paper proposes a novel algorithm BCC (Bag_of_cancerous_cells) to select the most significant histopathological features from the well-differentiated thyroid cancers. Methodology of proposed system comprises upon three layers. In first layer data preparation have been done by using BMF (Bag of Malignant Features) where each nuclei is separated with its related micro-architectural components and behaviours. In second layer decision model has been constructed by using CNN (Convolutional Neural Network) classifier and to train the histopathological behaviours such like BCP (Bags of chromatin Paches) and BNP (Bags of Nuclei Patches). In final layer, performance evaluation is done. A total number of 4520 nuclei observations were trained to construct the decision models from which BCP (Bags of Chromatin Patches) consists upon the 2650 and BNP (Bags of Nuclei Patches) comprises upon 1870 instances. Best measured accuracy for BCP was recorded as 97.93% and BNP accuracy was measured as 97.86%.**

***Keywords*—Medical Image mining; Decision support system; Pre-process; DICOM; FNAB**

I. Introduction

Recently classification of histopathological images is one of the active research area(s) of machine learning. Prediction of human cancerous cells and tissues would assume to be one of the most significant problems because micro-architectural components are likely to be found with heterogynous malignant behaviors and doctors are facing lots of confusions during the diagnosis phase. For example a set of micro-architectural components for each cancer type (such as well differentiated, poorly differentiated, benign cancers and other cancers) have deviated chromatin distribution, heterogynous nuclei behaviors, varying evidences for acentric nucleus within the set of nuclei and so on. In order to resolve these problems, some of very nice medical CAD (Computer added diagnosis) systems have been seen in recent past i.e. [1], [2] and [3]. These proposed approaches, addresses the classification problems of medullary, papillary, follicular carcinomas but yet cancerous behaviors such as chromatin level distortion, diffuse nuclei deviations and other cancerous behaviors are not reported in literature. Since efficient classification and feature selection of malignant behaviors would provide more dynamic assistance to doctors during the diagnostic phases because DICOM (Digital Imaging and Communications in Medicine) images are heterogynous in nature, always found with different shapes, sizes and structures at micro-architectural levels which depends upon the stage of different tumors as shown in[Figure 1]. In-order to provide assistance to cytologists in early diagnosis of cancer and to classify the malignant behaviors, this paper propose a system so called "Intelligent system for detection of abnormalities in human cancerous cells and tissues", which provides in-depth hidden knowledge of nuclei behaviors by proposing a novel algorithm BCC (Bag_of_cancerous_cells) where each nuclei is separated with its micro-architectural components so called BCP (Bags of chromatin Patches) and BNP (Bags of Nuclei Patches).

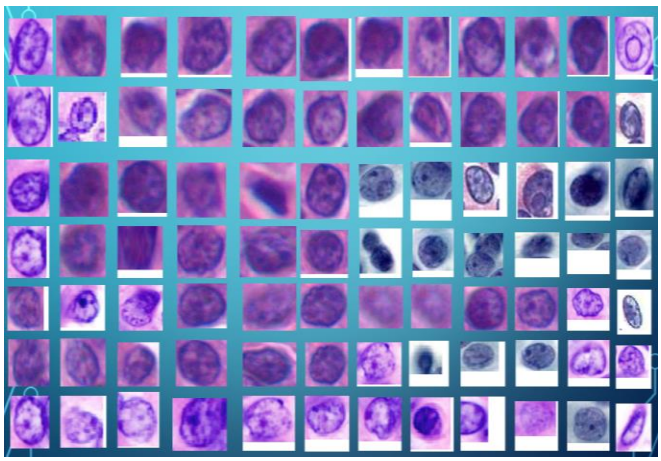


Fig. 1. Training Set of Nuclei with different behaviors- Normal Chromatin and abnormal Chromatin, Normal Nuclei and abnormal Nuclei with respect to appearance

The proposed system comprises upon the three layers, in first layer image pre-processing techniques have been used, in which noise reduction and feature selection is done by using proposed algorithm [algorithm 1] so called BCC (Bag_of_cancerous_cells). In second layer bags of chromatin patches and bags of nuclei patches are trained to construct the classification model based upon the deep learning algorithm such as convolutional Neural Networks.

A total number of 4520 nuclei were trained to construct the decision models from which BCP (Bags of chromatin Patches)

consists upon the 2650 and BNP (Bags of Nuclei Patches) was 1870 instances. Best accuracies for BCP (Bags of Chromatin Patches) and BNP (Bags of Nuclei Patches) were measured respectively as 97.93% , 97.86% .

Rest of paper is organized into the five sections, where section one is used to define the introduction and the second section reports the related works. Methodology is described in detail in section number three and the results have been presented into the section number four. Conclusion and discussion have been discussed into section five

II. Related Works

Basically this paper offers a productive modelling approach which deals with the classification problem of thyroid malignant diseases. Since histopathological DICOM (Digital Communication in Medicine) images needs special consideration to construct a decision model and to predict the micro-archetechral component behaviors such as abnormal chromatin distribution and heterogeneity between the set of nuclei. Many research approaches were proposed to solve the classification problem of cancerous cells and a few of them are presented as bellow.

A comparison [1] of different nuclei segmentation algorithms was proposed for thyroid disease by using the image clustering algorithms i.e. K-means and watershed in first stage as unsupervised learning and in second stage a template matching strategy algorithm was used for classification model as supervised instance learning. The best accuracies for both techniques were recorded respectively in final layer, 72% and 87%. Since micro- architectural components are very difficult to classify because of heterogeneity in terms of shapes, sizes and behaviors but the proposed approach of this paper deals the histopathological images at more deepest levels and it provides more effective assistance to doctors during the experimental setups to predict the interrelated properties of tumors at early stage.

A Comparison [2] of three ML (Machine Learning) neural network based algorithms was conducted to deeply analyze thyroid disease datasets. the classification model was built by using Scaled Conjugate Gradient, quasi-Newton method, Gradient Descent with Momentum and Bayesian regularization algorithms where gradient based layered features were used and the calculated best accuracies were approximated respectively as 90.5%, 86.30% and 83.50%. since the prediction of abnormal behaviors of cells is an essential problem at early stage but diffuse shape of pixels does not allow to select the appropriate set of pixels belonging to the features of regions of interest where cancerous material is persisting in DICOM image. The proposed algorithm of this paper auto detects and segments the cancerous regions by selecting the chromatin and nuclei behavior based feature. Since a dynamic threshold segmentation would allow to doctors to detect various regions of medical images at more granular levels because fixed threshold settings and intensities may not allow to detect a proper set of related attributes, there are maximum fare chances for loss of valuable image information but proposed approach of this article deals effectively with in-depth medical image segmentation and provides more precise assistance to doctors.

A system [4] was proposed for thyroid disease diagnosis and MIL (Multiple Instance Learning) was used as machine learning algorithm to predict the disease. Fully connected neuron model was constructed to classify cancerous thyroid disease tissues and best accuracy of the system was measured as 95.40%. Since the appropriate feature selection would reduce the computational complexity of CNNs (Convolutional Neural Networks) because effective algorithms would enhance the performance evaluation of a classifier. This paper contributes following three contributions.

A. *In literature [Table 5] previously fullicular, papilary, medullary cacer classification systems were reported but cancer behaviour classification is yet not reported. This paper offers a preductive modeling for the classification of above stated cancerious behaviours so called BCP (Bags of Cancer Patches) and BNP (Bags of Ncueli Patches).*

B. *Proposed segmentation algorithm BCC (Bag_of_cancerous_cells) provides more in-depth assistance for nuclei as well as chromatin detection using optimized segmentation technique and supports to build an efficient classification model by using CNNs into two distinct categories of thyroid cancer malignent behaviours.*

C. *The best measured ccuracies for both cancerious behaviours are calculated as BCP (Bags of Cancer Patches) 97.93% and BNP (Bags of Ncueli Patches) 97.86%.*

This paper uses state of art classification algorithm such as CNN (Convolutional Neural Networks) and our preprocessing algorithm reduces the complexity of pixel layers and every behavior is represented into 28 X 28 pixel size whereas the size of DICOM image is very high and needs significant time and memory constraints.

III. Methodology

The methodology of this paper falls into the category of machine learning and offers predictive modelling approach for classification of the biological behaviors of thyroid cancerous cells and tissues. This paper uses a real-world dataset of DICOM (Digital Communication in Medicine) images for FNAB (Fine Needle Aspiration Biopsy) received from Cytological department of affiliated hospital in Pakistan. Methodology of proposed system comprises upon three major layers as shown in [Figure 2]. In first layer datasets are prepared by using BCC (Bag_of_cancerous_cells). In second layer decision model is constructed by using CNN (Convolutional Neural Network) classifier by training the selected features from the malignant bags as BCP (Bags of chromatin Patches) and BNP (Bags of Nuclei Patches) to classify the abnormal behaviors of nuclei. In final layer performance evaluation is performed.

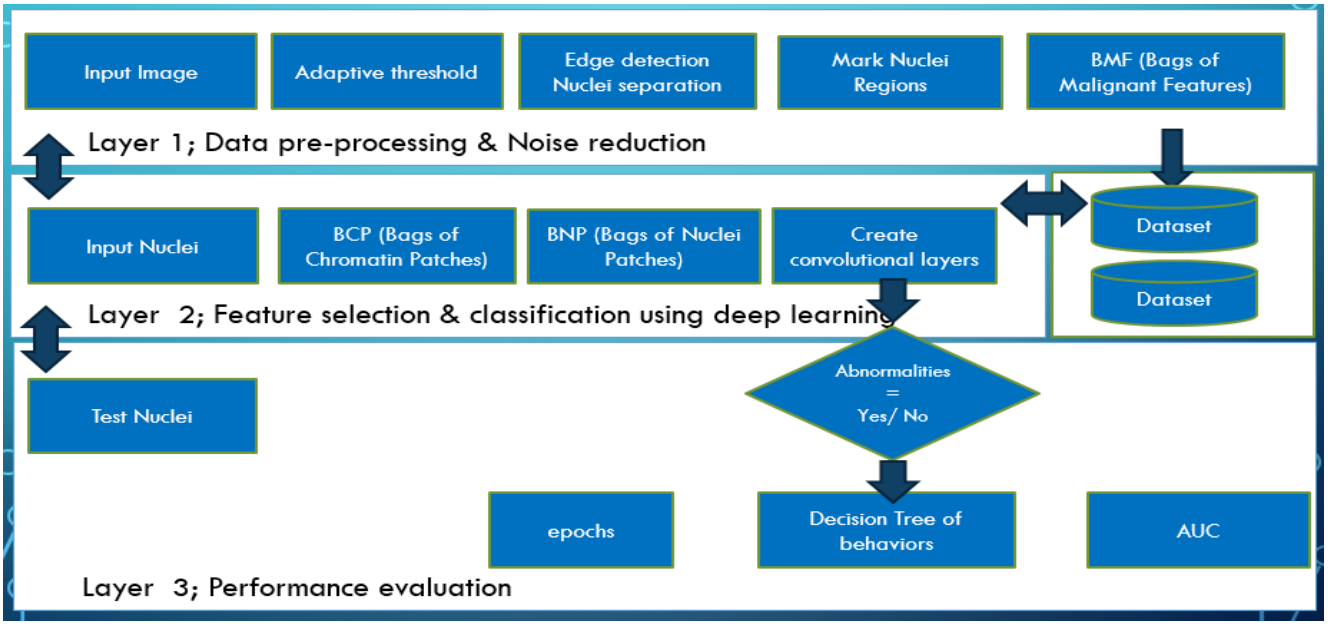


Fig. 2. Intelligent system for detection of abnormalities in human cancerous cells and tissues Workflow

A. *Layer 1: Data Pre-Processing & Noise Reduction*

Dataset: Due to un-availability of histopathological datasets of FNAB (Fine Needle Aspiration Biopsy) in literature, real-world datasets were prepared for training and testing purposes. Classification of abnormal behaviors of nuclei were performed at the deepest levels by using the selection of chromatin distribution and other nuclei related features which are not only providing meticulous assistance to doctors in quantification of micro-archetechral components but also helps to reduce the chances of misdiagnosis.

- *Noise Reduction:*

Noise reduction of DICOM (Digital Communication in Medicine) images is done by using adaptive threshold segmentation. Proposed pre-processing algorithm is presented in detail in following section as [algorithm 1] and [Figure 3].

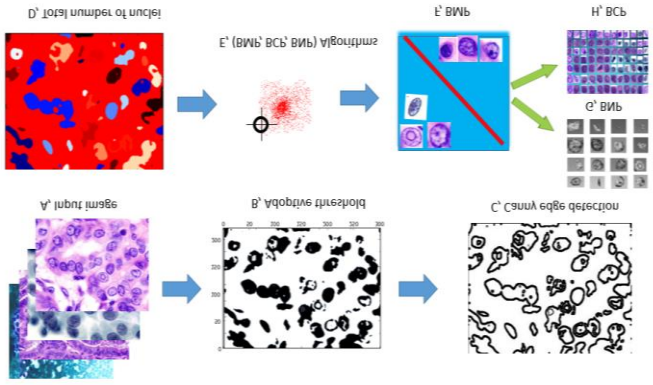


Fig. 3. Noise reduction and object detection

- *ALGORITHM 1:*

Let's consider an image consist upon the set of heterogynous attributes known as nuclei. Every nuclei in a set of image has its own behaviors and features. Firstly noise reduction was done and unnecessary information was eliminated, since the behaviors of nuclei are likely to be detected by using Otsu's global threshold method [Figure 3, B], where grey level intensity is considered as L means (0, 1… L-1) since L is assigned to approximate two classes by arranging the ranges 0 and 1. Those pixels who have the value of one are the pixels which represents some objects such as nuclei and zero range pixels are to be subtracted from the image by considering the noise. Let's consider $M^*N = n_0 + n_1 + \cdots \ldots \ldots \ldots + n_{L-1}$ are assigned to represent every pixel of the image, where n_i pixels are counted by i intensities, where two level thresholds represented by$T(k) = k, 0 < k < L-1$. Let's formulate c_1 as class$[0, k] and c_2 : [k+1, L-1]$.

$$P_1(k) = \sum_{i=i}^{k} p_i \quad (1)$$

$$P_2(k) = \sum_{i=k-1}^{L-1} p_i = P - 1_1(k) \quad (2)$$

Where each pixel P having the rang 1 such that$P_1 m_1 + P_2 m_2 = m_G, P_1 + P_2 = 1$, is the summarized measure of $P_1 + P_2 = 1$ and the global variance may be calculated by taking the slandered division of all pixels eq. (3)

$$\sigma_G^2 = \sum_{i=0}^{L-1} (i - m_G)^2 P_i \quad (3)$$

Class 0 intensity pixels may be acquired by using the slandered division of white pixels whose intensities would be considered $\sigma_B^2(k) = P_1 P_2 (i - m_G)^2$ where class 1 qualifying pixels would be divided to acquire the white matter of the objects by substation process of Otsu's method and adaptive threshold would represented as shown in eq.(4).

$$g(x,y) = \begin{cases} 1 & if\ (x,y) \geq k^* \\ 0 & if\ (x,y) \leq k^* \end{cases} \quad (4)$$

Algorithm 1: BCC (Bag_of_cancerous_cells)

Input: DICOM dataset containing set of Nuclei as D
$D \leftarrow HCV\ Nuclei\ components\ for\ Binarization\ B$
$Visit = each\ pixel\ as\ C\ (x_i, y_i)$
  $for\ each\ c_i \in T\ do$
  $T \leftarrow \frac{1}{2}(m_1 + m_2)$
  $C_j \leftarrow g(x,y) = \mu T$
  $P \leftarrow \frac{\sigma_b^2}{}(k) \leftarrow P_1 P_2 (m_1 - m_2)^2$
  $K \leftarrow g(x,y) \leq 0_n(k) \leq 1$
  $if\ g(x,y) = 1$
  $Count \leftarrow p_i\ 1 + +||Foreground||$
  $if\ g(x,y) = 0$
  $Count \leftarrow p_{i\ \vee} + +||Background||$
  $end\ if$
  $Return \leftarrow Foreground$
$Nuclei_{Edges} \leftarrow g_h(\tau) = h_b(x) * g(x)$
  $for\ each\ g_g(\tau) \in T\ do$
  $do\ h_b(x) = |b|^{[-]}\mathrm{csch}\ (\pi b x^{1.5})$
  HTL ← b = 0+ , b = ∞
  end
  $Return \leftarrow Nuclei_{Edges}$
$Homo_{Nuclei} \leftarrow T[b] = \{(s,t)|g(s,t) < n\}$
  $for\ each\ g(s,t)\ do$
  $n = min + 1\ to\ n = max + 1$
  $T = [n] = 0, otherwise\ 1$
  end
  $Return \leftarrow C[n] = \cup_{i=1}^{R} C_n(M_i), C_n(M_i)$
$Break_{Nuclei} \leftarrow \frac{\sum_{i=1}^{n} a(x_i)}{n} + \frac{\sum_{i=1}^{n} a(y_i)}{n}$
  $h, w \leftarrow \frac{\sum_{i=1}^{n} a(x_i)}{n} + \frac{\sum_{i=1}^{n} a(y_i)}{n} + +\mu h/w$
  $PCP =$ Bag of Chromatin Paches $\leftarrow Set\ of\ Nuclei\ seed(\mu x, y) \leftarrow Fourground$
  $BNP =$ Bag of Nuclei Paches $\leftarrow Set\ of\ Nuclei\ seed(\mu x, y) + \frac{\sum_{i=1}^{n} a(x_i)}{n} + \frac{\sum_{i=1}^{n} a(y_i)}{n}$
  $Return$ Bag_of_cancerous_cells $\leftarrow Set\ of\ Nuclei\ seed(\mu x, y)$
End

The edges [Figure 3] of each detected nuclei after segmentation process may be considered as $g_H(\tau) = h_b(x) * g(x), where h_b(x) = |b|\ \mathrm{csch}(\pi b x)$ where $g_H(\tau)$ has to find edge position of every set of nuclei where values of $h_b(x) = |b|\ \mathrm{csch}(\pi b x)$ have to qualify the threshold as approximated by eq. (5). A function eq.(6) have been created and the edges are detected by considering the HLT (High Level Threshold) at each connected corners of nuclei starts from zero to infinity eq(7).

$$G_H(f) = h_b(f) g(f)\ where\ G_H(f) = FT[g_{H(\tau)}] \quad (5)$$

$$G(f) = FT[g(r)], H_b(f) = \tan h\ (\pi\ f|b) \quad (6)$$

$$\text{HTL} = \text{b} = 0+ , \text{b} = \infty \quad (7)$$

Let's consider $g(s,t)$ is required intensity where s represents the number of empty nuclei and t is the threshold of every nuclei which is to be connected by colour scheme as per homogeneity and heterogeneity eq.(7) and eq.(8). We use watershed segmentation to find the these regions having $n = min + 1\ to\ n = max + 1$, where n is the number of connected components with min and max limits, since we count $C[n] = \cup_{i=1}^{R} C_n(M_i), C_n(M_i)$ where n is beneath number of T(n) which is randomly filled and counted as number of region in $\cup_{i=1}^{R} C_n$.

$$T[b] = \{(s,t)|g(s,t) < n\} \quad (8)$$

$$T = [n] = 0, otherwise\ 1 \quad (9)$$

Regional maxima library is used to extract the spatial attributes incorporated by Euclidian distances between the instances. Nuclei separation was done by radii cuts by considering the centroids of the image as spatial locations. As shown in [Figure 3] nuclei have been separated and converted into grey scale intensities because doctors use different staining material as stated above therefore in-order to absolve the effects of biomarkers since the grey-scale images have comparatively high accuracies in comparison with colour based nuclei during the classification layer. Let's consider a central locations with corner information $G(f) = FT[g(r)], H_b(f) = \tan h\ (\pi\ f|b)$ by considering each pixel, where h is the height of nuclei and w is the width of particular object as shown in eq(9). Since every separated nuclei has unique location in dataset D as $seed(\mu x, y)$eq.(10).

$$h, w = \sum_{a(y)=n}^{a(x)=1} + \mu h/w \quad (10)$$

$$seed(\mu x, y) = \{Sum(a(x,y)_{n+1}^{n-1})\} \quad (11)$$

On the basis of nuclei radii cuts were made and chromatin bags were prepared by considering the foreground of nuclei images and second data set of background was considered to classify the nuclei bags with the assistance.

B. Layer 2: Feature Selection & Classification Using Deep Learning

This paper presents classification of cancerous bags consisting upon the chromatin bags and abnormal nuclei bags. The CNN (Convolutional Neural Network) is a linear classifier since it uses weighted matrix W and bias vector b. Responsible to collect and forward input-values of image features to neurons (number of neurons or hidden layers depends upon the complexity of problem). Supplied data to neurons has to be calculated with some hidden 'bias' weights for further processing. Due to the neuron like structure one input is connected to another input vectors such like bias and each class of nuclei behaviors is represented by hyperplane by using the vector spaces. Let's consider vector$x \in class\ i$ and variable Y is a stochastic variable.

$$P(Y = i|x, W, b) = softmax_i(W_x + b) = \frac{e^{W_i x + b}}{\sum_j e^{W_j x + b_j}} \quad (12)$$

Where class Y_{pred} is the max probability to predict the model during the training layer.

$$y_{pred} = argmax_i P(Y = i|x, W, b) \quad (13)$$

Parameters supplied to train the classifier are responsible to maintain the state of persistence which are assigned to shared variable W, b in the parameters of $P(Y|_x, W, b)$ where x may be considered as cancer bags vector types. Since the optimal model for learning comprises our minimizing loss strategy but the more than one class (multi- class) strategy considers $-ve\ likelihood\ \mathcal{L}\ and\ loss\ \ell$ in a dataset D to follow the parameters as defined.

$$\mathcal{L}(\theta = \{W, b\}, \mathfrak{D}) = \sum_{i=0}^{|\mathfrak{D}|} log\left(P\left(Y = y^{(i)}\middle|x^{(i)}, W, b\right)\right)$$

$$\ell(\theta = \{W, b\}, \mathfrak{D}) -= \mathcal{L}(\theta = \{W, b\}, \mathfrak{D}) \quad (14)$$

Since the disease represented variables inputs are defined in x quantities with Y classes. In CNNs deep learning models all the training instances are fully connected to output layer. Gradient loss by considering the parameters as defined in $\partial\ell\ /\ \partial W$ and $\partial\ell\ /\ \partial b$ are able to handle a large number of classes but computational complexities takes huge training time because of low processing capabilities of normal desktop computers.

C. Layer 3: Performance Evaluation

The confusion matrix, Precision and recall measures are used to evaluate the performance of proposed systems as shown in [Table 1], [Table 2], [Table 3] and [Table 4]. Epochs

of classifier have been presented in [Figure 4], [Figure 5] where cumulative representation of instances have been shown by using CNN classifier. Since the pixel precision can be recalled by considering the probability retrieved variables with associated unknown inputs consisting upon unknown behaviors. The details of performance evaluation is defined in following result section.

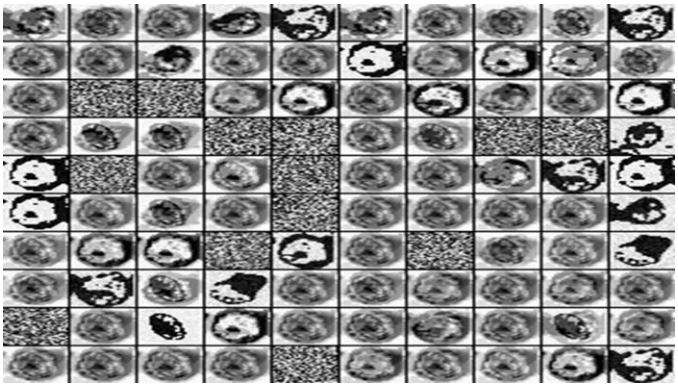


Fig. 4. Classifier epochs for Bags of Chromatin Patches- Behavior Chromatin distribution

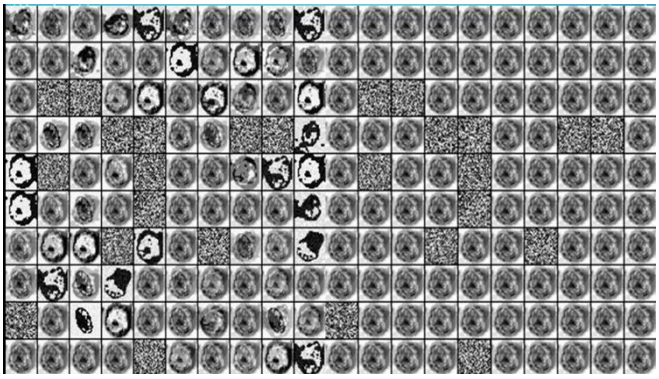


Fig. 5. Classifier epochs for Bags of Bags of Malignent Patches

IV. Results

A total number of 4520 nuclei were trained to construct the decision models. Confusion matrix [Table 1] for BCP (Bags of chromatin Patches) consists upon the 2650 nuclei. The results show that a number of 1566 observations were classified as true positive and 34 instances were miss-classified for cancerous class label attribute. The approximated precision and recall measure was recorded respectively 97.87% and 98.67%. For Non-Cancerous class label attribute 1039 observations were classified and 21 instances were remained miss classified with the approximated [Table 3] precision and recall measure as 98.01% and 96.83%. The measured classification accuracy for Bags of chromatin Patches was recorded as 97.93%. The confusion matrix [Table 2] for Bags of Nuclei Patches, 991 number of instances were classified for cancerous the class label attribute and 29 observations were miss classified with the precision and recall [Table 4] measure recorded about 97.15% and 98.90%. For Non-Cancerous class label attribute 839 instances were classified and 11 were determined as miss classified. Over all classification accuracy for Bags of Nuclei Patches was recorded as 97.86%. The cumulatively measured classification of the system is 97.91%, since the precision and recall measures were estimated about 98.70% and 96.56% respectively.

TABLE I. Confusion Matrix for BCP (Bags of Chromatin Patches)

Classifier	*BNF Cancerous*	*BNF Non-Cancerous*
BCP Cancerous	1566	34
BCP Non-Cancerous	21	1039
Classification accuracy		97.93%

TABLE II. Overall Performance of BCP (Bags of Chromatin Patches)

Raw BCP Images	No of Extracted Chromatin	No of Classified PCP	No of miss-classified BCP	Prece-ssion	Recall
2650	1600	1566	34	97.87%	98.67%
	1060	1039	21	98.01%	96.83%

TABLE III. Confusion Matrix for BNP (Bags of Nuclei Patches)

Classifier	*BCP Cancerous*	*BCP Non-Cancerous*
BNP Cancerous	991	29
BNP Non-Cancerous	11	839
Classification accuracy		97.86%

TABLE IV. Overall Performance of BNP (Bags of Nuclei Patches)

Raw BNP Images	No of Extracted Nuclei	No of Classified BNP	No of miss-classified BNP	Prece-ssion	Recall
1870	1020	991	29	97.15%	98.90%
	850	839	11	98.70%	96.65%

The estimated AUC (area under curve) is represented is [Figure 6], where AUC for BCP behavior is measure as 0.9385 and the AUC for BNP class was approximated as 0.9915. The comparison of this papers proposed approach with literature is presented in [Table 5], which shows that CNN classification produces the more enhanced accuracies.

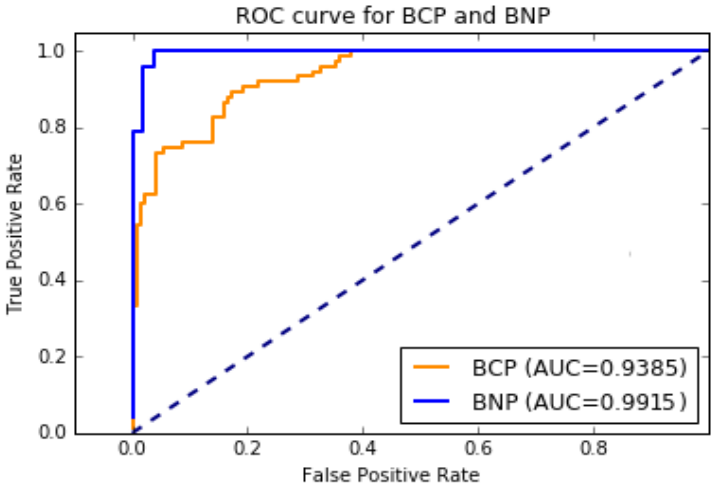


Fig. 6. AUC for behaviors BCP, BNP

TABLE V. COMPARISON OF OUR SYSTEM WITH LITERATURE

Approaches	*Image preprocessing Techniques*	*Machine Learning Technique*	*accuracy*
Cheng Chen, 2013	K-means	supervised learning-based template matching	72.00%
	watershed		87.00%
Pourahmad, 2012	Gradient based features were taken on 5, 10 and 20 neurons	Scaled Conjugate Gradient and BFGS quasi-Newton	90.50%
		Gradient Descent with Momentum	86.30%
		Bayesian regularization	83.50%
Xu, Y, 2014	Multiple Instance Learning	DNN Deep Neural Networks	95.40%
Proposed approach of this paper. (a-Bag of chromatin patches) (b-Bag of Nuclei Patches)	Gradient based features containing 3 layers of pixels and acquired through novel algorithm BMF	Convolutional Neural network (Decision Tree)	97.93%
		Convolutional Neural network (Decision Tree)	97.86%

V. CONCLUSION AND CONCLUSION

Machine learning techniques are playing key role in medical diagnosis and classification problem of histopathological images is one of the major problems. In this article thyroid classification problem is presented as use case to describe the proposed. Since the differentiation between the abnormal human cancerous tissues requires special techniques to preprocess because improper segmentation would become the cause of loss of valuable information related to chromatin and set of nuclei. This paper proposes a system so called "Intelligent system for detection of abnormalities in human cancerous cells and tissues", which provides in-depth nuclei behaviors classification by proposing novel algorithms BCC (Bag_of_cancerous_cells) where each nuclei is separated with its micro-architectural components i.e. BCP (Bags of Chromatin Patches) and BNP (Bags of Nuclei Patches). Proposed algorithm not only reduces the complexity of classifier by detecting the objects as BMF (Bag of Malignant Features) but also assists the doctors to classify cancerous and non-cancerous quantities such as Bags of chromatin Patches and Bags of Nuclei Patches. Cconvolutional Neural Networks were used to construct the classification models for both DICOM behaviors. A total number of 4520 nuclei were trained where [Table 1] BCP (Bags of Chromatin Patches) consists upon the 2650 nuclei observations and measured classification accuracy for BCP was 97.93%. In confusion matrix [Table 2] represented for BNP (Bags of Nuclei Patches) was measured as 97.86%. Additionally various kinds of cancers could be quantified by using proposed preprocessing algorithm to reduce the noise and to extract the appropriate features of interest. In future works this article recommends to resolve the classification problems of anaplast cancers which are most aggressive cancers occurs in human organs such like thyroid and ovary.

ACKNOWLEDGMENT

We are highly thankful to all officials of Sukkur IBA (Sukkur Institute of Business Administration) and medical partners for providing an opportunity to conduct the research in applied sciences.

REFERENCES

[1] Cheng Chen, Wei Wang, John A. Ozolek, and Gustavo K. Rohde., A flexible and robust approach for segmenting cell nuclei from 2D microscopy images using supervised learning and template matching. Journal of Cytometry A. 2013 May ; 83(5): p.p 495–507

[2] Pourahmad, S., Azad, M., Paydar, S., & Reza, H.. Prediction of malignancy in suspected thyroid tumour patients by three different methods of classification in data mining. In First International Conference on advanced information technologies and applications 2012. pp. 1-8.

[3] Y. Xu, T. Mo, Q. Feng, P. Zhong, M. Lai, E. I. Chang et al., "Deep learning of feature representation with multiple instance learning for medical image analysis", ICA SSP, 2014.

[4] Y. Chen, R. Ranftl, and T. Pock. Insights into analysis operator learning: A view from higher-order filter-based mrf model. IEEE Trans. on Image Processing, 2014

[5] Y.-W. Tai, S. Liu, M. S. Brown, and S. Lin. Super resolution using edge prior and single image detail synthesis. In CVPR, 2010.

[6] R. Timofte, V. De, and L. V. Gool. Anchored neighborhood regression for fast example-based super-resolution. In ICCV, 2013.

[7] J. Yang, J. Wright, T. Huang, and Y. Ma. Image super resolution as sparse representation of raw image patches. In CVPR, 2008.

[8] Zhi-Yong Wang and Zhenbiao Yang (eds.), Plant Signalling Networks: Methods and Protocols, Methods in Molecular Biology, vol. 876, Springer Science Business Media, LLC 2012.

[9] Stegmaier J, Otte JC, Kobitski A, Bartschat A, Garcia A, et al. Fast Segmentation of Stained Nuclei in Terabyte-Scale, Time Resolved 3D Microscopy Image Stacks. PLoS ONE 9(2): e90036, 2014.

[10] Wienert S, Heim D, Kotani M, Lindequist B, Stenzinger A, et al. CognitionMaster: An object-based image analysis framework. Diagnostic Pathology, 2013, p.p 8: 34.

[11] Fan, Y., Shi, L., Liu, Q., Dong, R., Zhang, Q., Yang, S & Wang, J. . Discovery and identification of potential biomarkers of papillary thyroid carcinoma. Mol Cancer, 8(1), 2009, p.p 79-93.

[12] Dimitris K. Iakovidis, L.. Fusion of fuzzy statistical distributions for classification of thyroid ultrasound patterns. Journal of Artificial Intelligence in Medicine 50 (2010) p.p 33–41.

[13] Ding, J., Cheng, H., Ning, C., Huang, J., & Zhang, Y. . Quantitative measurement for thyroid cancer characterization based on elastography.Journal of Ultrasound in Medicine, 30(9), 2011, p.p 1259-1266.

[14] Iakovidis, D. K., Keramidas, E. G., & Maroulis, D. (2010).. Fusion of fuzzy statistical distributions for classification of thyroid ultrasound patterns. Journal of Artificial Intelligence in Medicine 50 (2010) p.p 33–41.

[15] Bell A.A, Kaftan, J.N, Schneider, T. E... Imaging and Image Processing for Early Cancer Diagnosis on Cytopathological Microscopy Images towards Fully Automatic AgNOR Detection, WrithArchin science journal 2006.

[16] Dimitris G., Panagiota S., Stavros T., Giannis K., Nikos D., George N., Dionisis C, Unsupervised segmentation of fine needle aspiration nuclei images of thyroid cancer using a support vector machine clustering methodology, 1st IC-SCCE Athens, 8-10 September, 2004.

[17] Suzuki, H., Saita, S., Kubo, M., Kawata, Y., Niki, N., Nishitani, H.,& Moriyama, N. An automated distinction of DICOM images for lung cancer CAD system. In SPIE Medical Imaging (pp. 72640Z-72640Z). International Society for Optics and Photonics, 2009.

[18] Denkert, Carsten, et al. "Tumor-infiltrating lymphocytes and response to neoadjuvant chemotherapy with or without carboplatin in human epidermal growth factor receptor 2–positive and triple-negative primary breast cancers."Journal of Clinical Oncology 33.9 (2015): 983-991.

[19] Wienert, Stephan, et al. "CognitionMaster: an object-based image analysis framework." Diagnostic pathology 8.1 (2013): 1.

[20] Wienert, Stephan, et al. "Detection and Segmentation of Cell Nuclei in Virtual Microscopy Images: A Minimum-Model Approach." (2012).

[21] Kurman, Robert J., and Diane Solomon. The Bethesda System for reporting cervical/vaginal cytologic diagnoses. Springer Science & Business Media, 1994.

[22] Petushi, Sokol, et al. "Large-scale computations on histology images reveal grade-differentiating parameters for breast cancer." BMC medical imaging6.1 (2006): 1.

[23] Krystosek, Alphonse, and Theodore T. Puck. "The spatial distribution of exposed nuclear DNA in normal, cancer, and reverse-transformed cells."Proceedings of the National Academy of Sciences 87.17 (1990): p.p 6560-6564.

[24] Gurcan, Metin N., et al. "Histopathological image analysis: A review." IEEE reviews in biomedical engineering 2 (2009): p.p.147-171.

[25] Cireşan, Dan C., et al. "Mitosis detection in breast cancer histology images with deep neural networks." International Conference on Medical Image Computing and Computer-assisted Intervention. Springer Berlin Heidelberg, 2013.

[26] Dalle, Jean-Romain, et al. "Nuclear pleomorphism scoring by selective cell nuclei detection." WACV. 2009.

[27] Doyle, Scott, et al. "Automated grading of breast cancer histopathology using spectral clustering with textural and architectural image features." 5th IEEE International Symposium on Biomedical Imaging: From Nano to Macro. IEEE, 2008.

[28] Mulrane, Laoighse, et al. "Automated image analysis in histopathology: a valuable tool in medical diagnostics." Expert review of molecular diagnostics8.6 (2008): p.p 707-725.

[29] Irshad, Humayun, et al. "Methods for nuclei detection, segmentation, and classification in digital histopathology: a review—current status and future potential." IEEE reviews in biomedical engineering 7 (2014): p.p 97-114.

Enhanced Re-Engineering Mechnanism to Improve the Efficiency of Software Re-Engineering

A. Cathreen Graciamary
Research Scholar, BhararthyiarUniversity, Coimbatore

Dr. Chidambaram
Asst prof, Rajah serofiji College, Thanjavur

***Abstract*—Generally, software re-engineering is economical and perfect way to provide much needed boost to a present software system. Software Re-engineering is like to obtain a fully completed software from existing software with additional features if needed. The overall process of Software re-engineering is to analyze the needed requiements & its contents. It also changes the needed contents or transforms the existing software system for reconstructing a novel software system. The difficult part in re-engineering is to understand the traditional system. Most of the software re-engineering mechanisms are aimed to achieve the common re-engineering objectives and the objectives are: improved software quality, reduced complexity, reduce maintenance cost and increased reliability. As a result, several traditional re-engineering mechanisms fail to verify the performance of individual functionality in existing software. This performance evaluation increases the complexity in re-engineering process. To minimizing the complexities in software re-engineering, this proposed system implements a novel approach named Enhanced Re-engineering mechanism. This enhanced mechanism introduces a new idea, before executing the re- build process the developer verifies the performance of particular function in existing system. After that, the function performance is compared with proposed algorithm. Based on the comparison process only rebuild process should be carried out. Finally this proposed mechanism reduces the complexities in software re-engineering.**

Keywords*—Software Engineering; Software Re engineering; Software Quality; Restructuring

I. INTRODUCTION

The reengineering process of the software is modifying and reorganizing the existing system of software for making them maintainable [1]. This is a part of rewriting or restructuring the entire system of legacy without functionality changes. The main intention of the reengineering process is to modify or modernize the previous system in a newer version. The changing process of the business becomes too difficult and complex, or too much costly for the implementation. The legacy maintenance is too costly than the solution of these issues [2]. The benefit of software re-engineering is reducing the risk level at low cost.

The process of re-engineering might face several risks like the software engineering is facing. The identification of the risk is a technique. The identification of the risk is important for the assessment of the risk, management and the analysis of the risk. In proposed system, the risks over the potential are being classified and analyzed [3]. The monitoring technique used to classify and analyzed the risk. It is helping the system of reengineering over the maintenance of ease and profit over cost with the low risk at low cost.

A. Re-engineering- An Overview

The process of reengineering generally includes the grouping of different process like forward engineering, re-documentation, reverse engineering and translation. The reengineering objective has been explained in the fig.1 that makes easy to realize the existing software or its original functionality, redesigning and enhancement with the latest technology, and added subsystem for gaining more benefit with current technology [4].

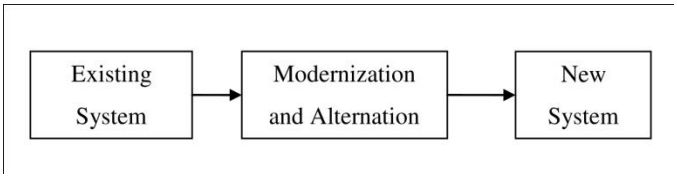


Fig. 1. Re-engineering Process

The latest system which is in use is known as the legacy system. When the system of legacy becomes weak or old for this design, coding, structure, and new module could be added for making the strong legacy system. This process is making the system hard and it is not going to be beneficial over the cost [5].

The software interpretation is simply large and it includes the record of design, other documentation sources and source codes. The reengineering of software is being portioned into two activities sets.

1) The first activity set containing the understanding of supporting program, reverse re-engineering, browsing and measurement.

2) The second activity set containing the geared evolution of software, like re-modularization, re-structuring, and re-documentation.

B. Forward Engineering

Forward engineering process is initially begins from requirement specification phase to the software implementation phase. Forward engineering is also known as reclamation or renovation; this process not only extracts design information from traditional software application, but it uses design information to reconstitute or alter the traditional software system in an attempt to get better and overall quality of software application. In most of the cases, re-engineered software application re-executing the existing system

functions and also includes a new functions to improve total performance.

C. Reverse Engineering

This process initially begins from the software implementation phase then moving in the direction of developing, designing and requirement specification phase [6]. Thus, procedural design and architectural information of data was discovered from traditional legacy software application.

D. Strategic Reengineering

It refers to the re-development of software system for meeting the criteria of long term plan of the company standard [7]. The lifecycle of the strategic reengineering are involved in the four phases that is planning of Information System (IS), planning reengineering, software reengineering and the building of the reusability framework.

E. Definition of software reuse

The definition of the software reuse is mentioned below [8]:

1) "The systematic procedure for creation or development of the software from the stock of blocks, so, the requirement similarity or the architecture in amid of the application could be exploited for achieving the benefit in business performance, productivity and quality".

2) "The capability for using the routines of software in novel application".

F. Why Reuse Software?

The reuse of a better or efficient software provide the facility to increase the reliability, quality and productivity and reducing the time of implementation and cost of the software [9]. The repository and reuse process of the software is producing the knowledge base, which is improving the quality of software after each process of reuse, reducing the size of developing work that require for the projects of future and minimizing the novel project risk that based on the knowledge repository.

For that purpose, this proposed system implements a novel approach named enhanced re-engineering mechanism for minimizing the complexities in software re-engineering. In this mechanism, this system compares the performance of functionalities of existing software with the functionalities of new software. Based on the performance evaluation process, re-build process can be carried out.

The rest of this research work is organized as follows. Section 2 states the related work of this proposed work. Section 3 presents the detailed description of this proposed work. Section 4 presents the results and discussion part. Section 5 concludes the future work of this proposed work.

II. Related Work

Normally Re-engineering process discuss about the "alteration of business process". The changes over the process of business need some novel requirement over the systems [10]. The researcher in [12] has included the changes in process of business, even the present situation for occurring problem also with over time in the development of a system in any organization and that have to be utilized in another organization. There are no more expert in the field of reengineering like design field and in those reengineering, engineers also don't have experience over the research field. Legacy system problem had posed all over the place in world. The Research work [13] is defining about the Legacy systems that significantly oppose evolution and modification of the new changes over the requirement of the business in spite of using technology for designing. The system of legacy is being replaced with the some novel system with the improved and same functionality [11]. The author of the [14] has mentioned in their research work that presenting the reengineering of the iterative over the function of legacy that describe the gradual reengineering process of the components procedural legacy system. The proposed technique of this research work is enabling the system of legacy for making gradual empty over the system of reengineering without freezing the legacy system or duplicate system of legacy. This process is containing the components of legacy system that initially restore the system and then move towards the system of reengineering. At the same time, legacy system could exist in both parts reengineering and restoring of the system. At the last stage of the process, a single system could be existed that is reengineered system. The technique has applied over the original system of reengineered and displayed the ability of it that is support for the gradual reengineering, maintenance of the system during the process of work, the request for the minimum requirement for the freeze maintenance, renewal of the reengineered system for operative environment within the system of legacy and, eliminate every system over the symptoms of aging. The reengineering model of Dual-Spiral for proposed legacy system research work [15] performs a cyclic approach. The reengineering model of the Dual-Spiral workflow require two system together for the work (Each target and legacy requires one system respectively), and functionality move from legacy to target stepwise. During the whole process, the legacy system active functionality is in the pattern of decree-mental, and the novel system of the target over the active functionality is in the pattern of incremental.

For improving the Legacy system quality, the process of software re-engineering must have to enable its new technologies and functions to make sure the efficiency of information management contain the legacy system. The reengineering of Software process involves in the re-documenting and restructuring of legacy system through adding the evolution attempt for easy maintenance and increase the competition and rises over the technology that have forced the organization to adopt new and enhanced approaches for the renovation of the system of legacy within the product, service and process. The method of the evolution over the reengineering of software helps to gain the control over the effective cost, improvement in the quality and the reduction over the risk and time. In few last years some of the framework for reengineering and risk management was developed, but only some of framework was able to recognize the factor of risk in the process of reengineering for successfully creating the risk solution of reengineering by system of software. The research work [18] has developed metrics framework for evaluation of the software system

legacy complexity for outsource support. Legacy system framework considers the two dimensions: source code and documentation.

The author in [16] has described the gradual reengineering process over the legacy system for the procedural components. The proposed technique enables the system of legacy to become gradually empty in the system of reengineering; there is no need to do freeze and duplication of legacy system. The research work [19] has focused on the reengineering aspects of technical and the political risk; these are the reason of reengineering effort failure. But, there is more other risks are available in it like risk of reliability, risk of performance, risk of technology, risk of complexity, risk of availability, risk of security, risk of modularity, and risk of usability [17]. Re-engineering risk measurement and identification is necessary competence for successful effort of software re- engineering that provides better strategies for the improvements of quality, control over the cost, and reducing the risk and time for the evolution of legacy system.

III. PROPOSED SYSTEM

A. Proposed Work Overview

Enhanced re-engineering mechanism is a software re-engineering process that exploits not just a single, but a mixture of abstraction methods and abstraction levels to transform a traditional software system to a new software system. The figure.1 illustrates the overall process of proposed Enhanced Software re-engineering mechanism. Here, proposed system utilizes both forward engineering and reverse engineering techniques. Initially, this proposed mechanism performs feasibility study to check the compatibility of system, and then analyze the components required for that re-engineering process. After collecting the requirements, it move to the next phase. In 2nd phase, mapping of Restructured Software Requirements Specification (SRS) to complete the design document i.e. integration of new Software Requirements Specification to obtain the redesigned document. This re-designed document is an output of the 2nd phase. After completing phase 2, the process is moved to the next phase. In phase 3, the programming part is being customized based on the changes done in the re-design document. It can able to backtrack from this phase to second phase and vice versa if it is required. Then this proposed mechanism perform re-testing and re-integration of various software modules to execute a particular functionality. In this phase, proposed system compares the performance of functionalities of existing software with the functionalities of new software. As per the performance result, better algorithm is replaced with the existing system. This process reduces the complexity and improves the quality of new re-engineering software. After the completion of various modules integration, there is a need to implement the modified system and get the target system which is required by the user.

B. Working Methodology

This Enhanced re-engineering mechanism consists of five important phases. These are

1) *Feasibility study and requirements*
2) *Restructured System Requirements Specification*
3) *design to code*
4) *Comparison of Existing and proposed Functionalities*
5) *Implementation*

1) Feasibility study and requirements

In this proposed work the initial stage of Enhanced Re-engineering mechanism is the feasibility study and requirements. In this stage, the feasibility study of re-engineering is done i.e. verify the configuration and compatibility of the computer system. After completing the feasibility study, the needs are re-specified based on the user's demand. The SRS contains the entire requirements in a written structure and is an authorized document. To re-specify the system requirements, this system need to map this with the Software Requirements Specification.

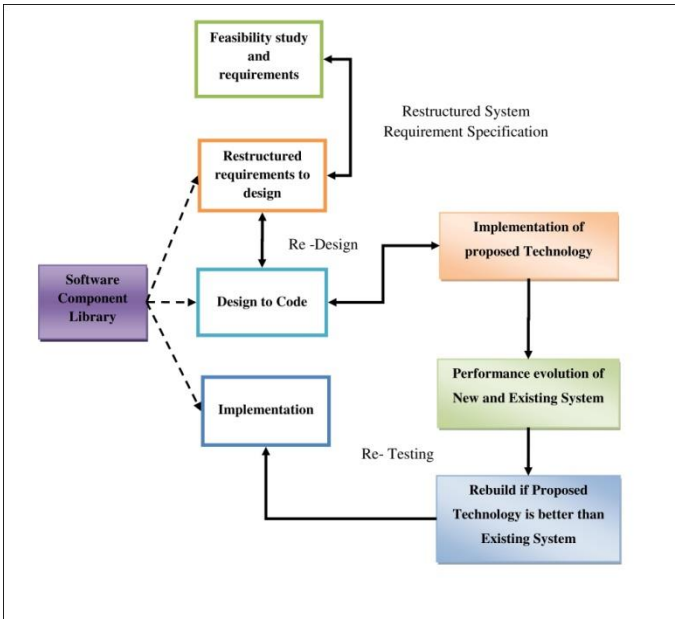


Fig. 2. System Architecture

2) Restructured System Requirements Specification

This stage describes in detail about the restructured Software Requirements Specification Process. Documentation is a significant attribute in the software development process as it reproduces the entire components of the total re-engineering process and performs as a blue print for the end product. Here, the experts compare the requirements of existing system with new proposed mechanism. SRS is used to integrate the new Software Requirements Specification with existing Software Requirements specification.

3) Design to code

This stage provides the details about design to code process. In this stage, as per the re-design document code has been done by the programmer. Usually, legacy algorithms are implemented in traditional development languages with existing functionalities which are required to be re-written in new functionalities. For example, this system has created one diagnosis system with Naïve Bayes algorithm. But as per the technology requirement the necessary of time and accuracy improvement is very important. So it has planned to re-engineer this application with SVM techniques.

4) Comparison of Existing and proposed Functionalities

This stage provides the details about Re-testing process. For Re-testing, initially takes both existing software application and a new software application. Then, it compares the performance of functionalities of traditional software application with the functionalities of new software application. For performance evaluation, this system utilize the metrics like running time, memory usage and system configuration. Then, the existing function performance is compared with proposed algorithm. Based on the comparison process, rebuild process should be performed. In comparison result, if an algorithm obtains more performance than other algorithms then the better performance algorithm is taken for re-build process.

5) Implementation

This stage is the final stage of Enhanced Re-engineering mechanism. As per the result of previous re-engineering stages, the implementation of a software application can be carried out. In implementation, a specific part can replaced with the good one which fully depends on the previous four stages of this Enhanced Software Re-engineering mechanism.

IV. Results and Discussion

A. Experimental setup

In order to analyze the performance of this proposed Re-engineering mechanism a series of experiments on a below mentioned dataset were conducted. In these experiments, this system implemented and evaluated the proposed methods in following configuration: Intel i3(R), CPU G2020, 2GB RAM, processer speed 2.90 GHz, operating system -Windows 7, Front End -JAVA and Back End-MySQL.

B. Dataset Details

TABLE I. Details of Datasets

S.No	***Dataset Name***	***Attributes***	***Class***
1.	breast-cancer	09	02
2.	Diabetes	08	02
3.	Heart Disease	13	02

For Brest Cancer,

Attribute 1: Clump_Thickness integer [1,10]
Attribute 2: Cell_Size_Uniformity integer [1,10]
Attribute 3: Cell_Shape_Uniformity integer [1,10]
Attribute 4: Marginal_Adhesion integer [1,10]
Attribute 5: Single_Epi_Cell_Size integer [1,10]
Attribute 6: Bare_Nuclei integer [1,10]
Attribute 7: Bland_Chromatin integer [1,10]
Attribute 8: Normal_Nucleoli integer [1,10]
Attribute 9: Mitoses integer [1,10]
Attribute Class {benign, malignant}

For Diabetes,

1. Number of times pregnant
2. Plasma glucose concentration a 2 hours in an oral glucose tolerance test
3. Diastolic blood pressure (mm Hg)
4. Triceps skin fold thickness (mm)
5. 2-Hour serum insulin (mu U/ml)
6. Body mass index (weight in kg/(height in m)^2)
7. Diabetes pedigree function
8. Age (years)
9. Class variable (0 or 1)

For Heart disease,

1. age
2. sex
3. chest pain type (4 values)
4. resting blood pressure
5. serum cholesterol in mg/dl
6. fasting blood sugar > 120 mg/dl
7. resting electrocardiographic results (values 0,1,2)
8. maximum heart rate achieved
9. exercise induced angina
10. old peak = ST depression induced by exercise relative to rest
11. the slope of the peak exercise ST segment
12. number of major vessels (0-3) colored by fluoroscopy
13. thal: 3 = normal; 6 = fixed defect; 7 = reversible defect
14. Absence (1) or presence (2) of heart disease

C. Performance Evaluation

In this proposed system, it has created one medical diagnosis and prediction system for three different diseases like heart disease, breast cancer and diabeties patients. Here it consider two medical diagnosis and predication applications such as

- Naive Bayes Classification System
- SVM Classification System

At first, this system evaluate the performacne of Naive Bayes Classification System. In this, three types of datasets are taken as an input for classification proeess. This classification system works based on the probability distribution of each data set and it efficiently predict the diseases. This predication fully depends on the probability distribution. But computation time of Naive Bayes classification system is too high. Later, it evaluate the performacne of Support Vecor Mechine Classification System and the same three types of disease datasets are taken as an input for classification process. This SVM classification system classifies the data sets based on the training process. This system also effectively predict the datasets,but the Computation time of this algorithm is very low.

Here comparison performance of both Naive Bayes Classification System and SVM Classification System showed. Based on the performance, SVM classification System is better than Naïve Bayes Classification System because the time consumption and prediction accuracy of Support Vector Machine is very High. Inorder to improve the Naive Bayes Classification System, software Re-engineering is required. In this proposed Re-engineering mechanism, there is no need to re-engineering the entire classification system. Instead of Naive Bayes Classification Algortithm, this system replacing the Support Vector Machine. So, the performance of existing classification system is also increased. This proposed

re-engineering process reduce the complexities in software re-engineerring process.

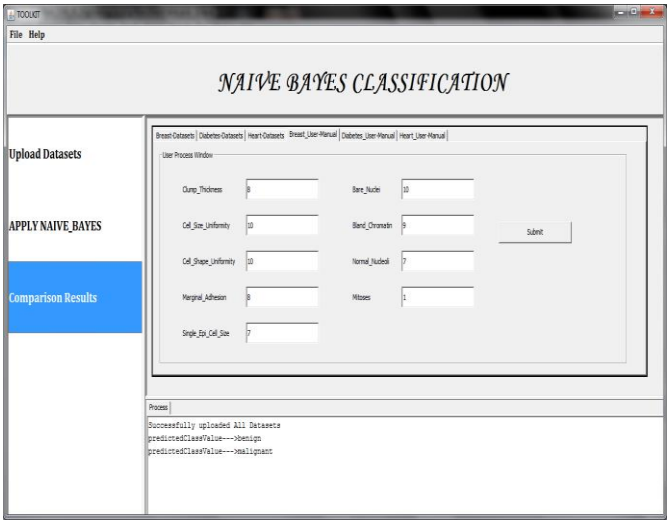


Fig. 3. Naive Bayes Classification System

The above figure represents a medical diagnosis and prediction system called Naive Bayes Classification system for predicting the various types of diseases like breast cancer, heart disease and diabetes.

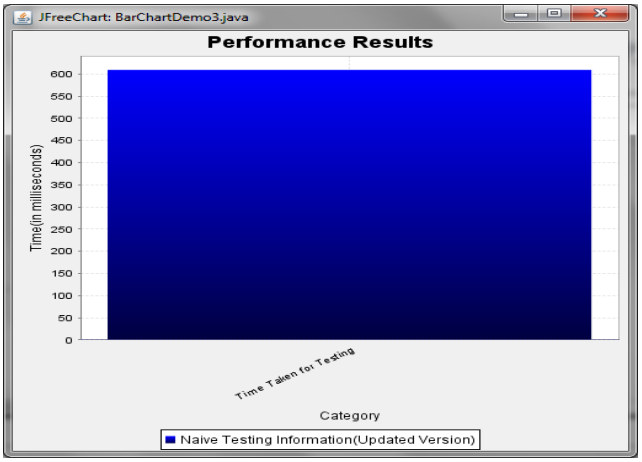


Fig. 4. Performance Evaluation of Naive Bayes

Classification system

The above graph represents the performance evaluation for Naive Bayes Classification system. For performance evaluation it consider execution time parameter and the results shows that it takes maximum time for prediction.

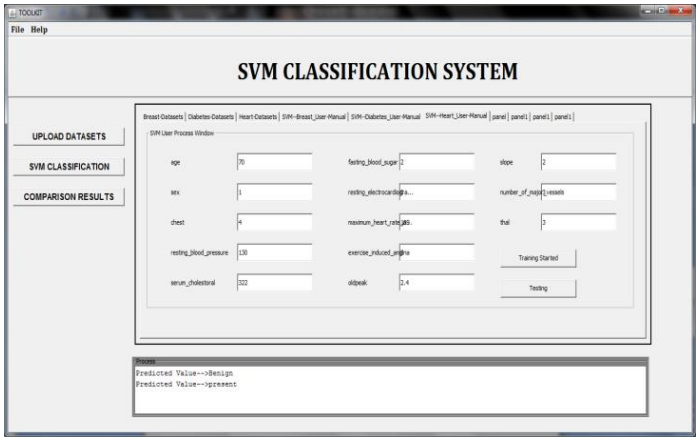


Fig. 5. SVM Classification System

The above figure represents a proposed medical diagnosis and prediction system called Support Vector Machine Classification system for predicting the different types of diseases like heart disease, diabetes and breast cancer.

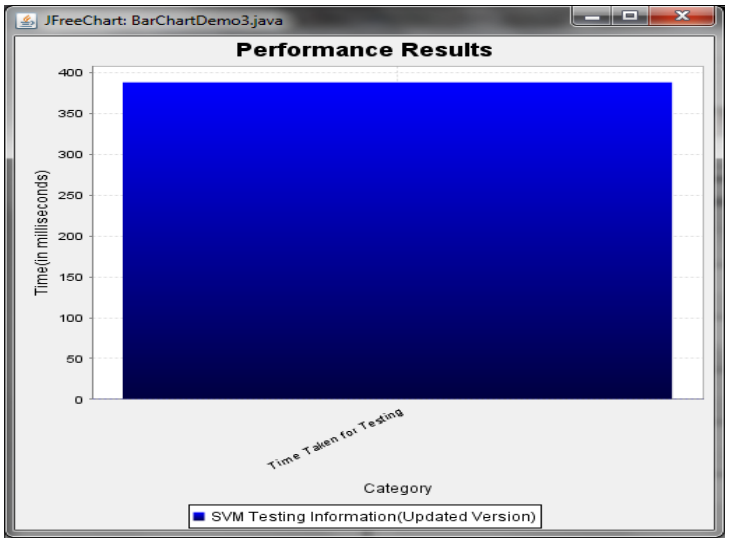


Fig. 6. Performance Evaluation of SVM Classification System

The performance graph represents the time consumption process of SVM classification system. Here, the graph shows the proposed mechanism takes less time for prediction process.

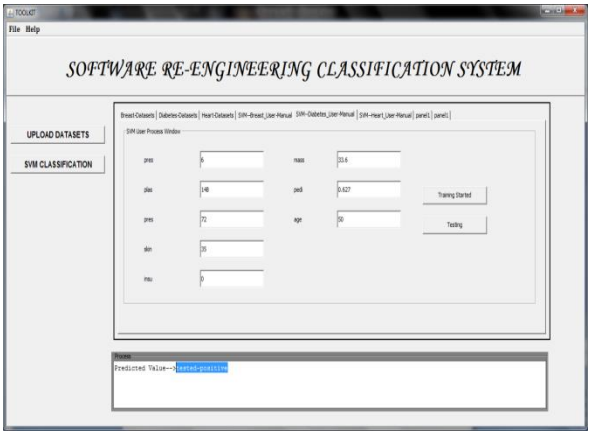


Fig. 7. Software Re-engineering Classification System

The above figure illustrates the proposed software reengineering classification system window. In software reengineering process, based on the performance evaluation result, the best algorithm is replaced with existing algorithm. Finally, the re-engineering process is successfully completed.

V. CONCLUSION

Nowadays, a lot of changes are rapidly involved in hardware and software because of the rapid growth of computer industry. Reengineering of software provides minimized risk level. New software development is really a high risk process because the software development process has some problems like development issues, specification problem, employee problems and cost. Software reengineering process overcomes the abovementioned software development problems because in re-engineering process some of the parts should be changed. To improve the performance of software re-engineering, our proposed system implements an enhanced reengineering mechanism. This mechanism is proposed to reduce the cost and time. Finally, our proposed system increases reliability of software and improve the quality of service with minimum development efforts.

REFERENCES

[1] Francisca O. Oladipo, Jude O. Raiyetumbi," Re-Engineering Legacy Data Migration Methodologies In Critical Sensitive Systems", Vol 6, No . 11, Journal of Global Research in Computer Science, November 2015.

[2] Manojkumar. P.K , " Classification And Analysis Of Risks In Reengineering Projects " , Vol . 2 , Nehru E- Journal - A Journal for Arts, Science and Humanitis , June –Dec 2014

[3] Hausi A. M¨ uller , Jens H. Jahnke , Dennis B. Smith , Margaret-Anne Storey , Scott R. Tilley , Kenny Wong , "Reverse Engineering : A Roadmap " , ICSE , 2000.

[4] Frank Weil, Ph.D , UniqueSoft, LLC , " Legacy Software Reengineering " , UniqueSoft LLC , 2015.

[5] Dr. Larisa Melikhova, Albert Elcock, Andrey A. Dovzhikov, Georgii Bulatov, Dr. Dmitry 0. Vavilov, "Reengineering for System Requirements Reuse: Methodology and Use-Case", 2007, IEEE.

[6] Krishan Kumar , Prabhpreet Kaur , " A Generalized Process of Reverse Engineering in Software Protection & Security " Vol. 4 , International Journal of Computer Science and Mobile Computing , May 2015

[7] T.G.J. Schepers , M.E. Iacob , P.A.T. Van Eck, "A lifecycle approach to SOA governance "https://www.researchgate.net/publication/234810165 , January 2008.

[8] M.H.Arifa Banu, N.Mohamed Thoufeeque, K.Archana, "Study of Software Reusability in Software Components", Vol 5 No 3, International Journal of Engineering and Technology, Jun-Jul 2013.

[9] Mohammed-V Agdal Univ, Ecole Mohammadia d'Ingénieurs (EMI), Siweb Research Team , "All About Software Reusability: A Systematic Literature Review " , Vol.76. No.1 , Journal of Theoretical and Applied Information Technology , 10th June 2015

[10] Ahmed Saleem Abbas, W. Jeberson & V. V. Klinsega, "Implementation of Fusion Model to Re-Engineer Legacy Software", International Journal of Computer Science and Engineering (IJCSE), 2013.

[11] Shekhar Singh, Significant role of COTS to design Software Reengineering Patterns, International Conference on Software Engineering and Applications(ICSEA),2009.

[12] Daniel Gjörwell, Staffan Haglund, Daniel Sandell," Reengineering And Reengineering Patterns", The Department for Computer Science and Engineering Mälardalens Högskola, 2002-02-24.

[13] Key Management Group Inc., " Legacy System Data Conversion And Migration " , Legacy System Data Conversion and Migration.

[14] Napas Methakullawat and Yachai Limpiyakorn, " Reengineering Legacy Code with Model Transformation " , Vol.8, No.3 , International Journal of Software Engineering and Its Applications , 2014

[15] Xiaohu Yang et al, "A Dual-Spiral Reengineering Model for Legacy System", TENCON 2005 - 2005 IEEE Region 10 Conference ISBN: 0780393112 Year: 2005 Pages: 1-5 Provider: IEEE Publisher: IEEE.

[16] Alessandro Bianchi, Danilo Caivano,Vittorio Marengo, Giuseppe Visaggio," Iterative Reengineering of Legacy Functions", 17th IEEE International Conference on Software Maintenance (ICSM'01),Florence, Italy, ISBN: 0-7695-1189-9,November 07-November 09.

[17] Er. Anand Rajavat , Dr. (Mrs.) Vrinda Tokekar , "Techrisk –A Decisional Framework To Measure Technical Dimensions Of Legacy Application For Rejuvenation Through Reengineering " , Vol.2, No.3, International Journal of Software Engineering & Applications (IJSEA), July 2011

[18] Oliver Hummel , Stefan Burger , "A Pragmatic Means for Measuring the Complexity of Source Code Ensembles " , IEEE. , 2013

[19] Basem Y. Alkazemi, "A Framework To Assess Legacy Software Systems", Vol. 9, No. 1, Journal Of Software, January 2014.

Scalable Scientific Workflows Management System SWFMS

M. Abdul Rahman
Department of Computer Science, Sukkur Institute of Business Administration
Sukkur, Pakistan

***Abstract*—In today's electronic world conducting scientific experiments, especially in natural sciences domain, has become more and more challenging for domain scientists since "science" today has turned out to be more complex due to the two dimensional intricacy; one: assorted as well as complex computational (analytical) applications and two: increasingly large volume as well as heterogeneity of scientific data products processed by these applications. Furthermore, the involvement of increasingly large number of scientific instruments such as sensors and machines makes the scientific data management even more challenging since the data generated from such type of instruments are highly complex. To reduce the amount of complexities in conducting scientific experiments as much as possible, an integrated framework that transparently implements the conceptual separation between both the dimensions is direly needed. In order to facilitate scientific experiments 'workflow' technology has in recent years emerged in scientific disciplines like biology, bioinformatics, geology, environmental science, and eco-informatics. Much more research work has been done to develop the scientific workflow systems. However, our analysis over these existing systems shows that they lack a well-structured conceptual modeling methodology to deal with the two complex dimensions in a transparent manner. This paper presents a scientific workflow framework that properly addresses these two dimensional complexities in a proper manner.**

Keywords*—*Scientific Workflows; Workflow Management System; Reference Architecture

I. INTRODUCTION

Over the last few years, we have witnessed a dramatic change in the way science and engineering has been conducted. In particular, computation became an established third branch of the science alongside theory and experiment. Scientific experiments can be classified into two parts, i.e. dry-lab and wet-lab. Dry-lab refers to the experiments that are conducted through computer-supported and automated computational (analysis) pipelines such as workflows in silico chemistry. Whereas wet-lab experiments attempt to focus on carrying out the experiments that involve manual tasks and human agents, for instance setting up the machines and preparing as well as measuring the samples. However, we use 'scientific experiment' as a comprehensive term which encompasses both parts of the experiments.

As a matter of fact, the management of scientific experiments is two dimensional, i.e. Application Management and Data Management. The former dimension refers to the management of the tasks (work steps) that are specific to the application domain such as collecting, preparing and measuring the samples, setting up and using scientific machinery and equipment, and performing computations and analysis while latter dimension addresses the tasks (e.g. data provision and preparation) that are solely related to the management of the data products generated from these domain specific applications. In fact, today's complex computational / analytical applications (tools) and heterogeneity of the data raise the intricacy from both dimensions, making the scientific experiments more challenging for domain scientists. Likewise, the proliferation of data generating devices, such as plasma-mass spectrometer in the computational chemistry and sensors in the meteorological research, makes it even more difficult since the data stemming from such type of devices are mostly noisy, inconsistent, rapidly changing, highly heterogeneous and incomplete.

Obviously, handling the Application Management dimension is not a trouble-free and effortless task; however the most challenging dimension is Data Management where domain scientists are in fact uncomfortable. This is mainly due to the following key **factors**:

Objective and Interest: For domain scientists, the prime objective is to conduct experimental study in order to get analysis and observation results. Thus, they are comparatively more intended towards the experimental tasks (e.g. analysis, computation and observation), rather than handling the data preparation tasks. Generally, the experimental specification is developed after a theoretical study of the domain. Thus it is believed that the scientists are the main community group that is assumed to be the responsible for designing and developing the experimental tasks. Moreover, our real world experience in diverse scientific domains demonstrates that the domain scientists do not seem to be so interested in specifying data management tasks; rather they get the assistance from data experts (other technological experts) for specifying and annotating these kind of tasks. This experience clearly concludes that scientific community group is rather more intended towards the management of the first dimension (Application Management) and shows less interest towards the management of the second dimension (Data Management).

Experience and Knowledge / Familiarity: This is common understanding that designing and developing the experimental steps (first dimension) require in-depth knowledge and experience of scientific domain while designing and developing the data management steps (second dimension) urges the in-depth and extensive knowledge about data related technologies. Scientists are assumed to be the community that has knowledge, experience and expertise about

their scientific domain and hence they better know their experimental tasks and the applications used in these tasks, for instance how to setup the scientific machines, what parameters should be set and how to prepare the samples for observation, how much and what kind of calibration sample should be used and so on. On the other hand, due to their insufficient experience and expertise towards data-related technologies they are experiencing comparatively more difficulties in defining data management tasks.

Due to the two dimensional complexity of scientific experiments, scientists are facing two types of key management challenges, i.e. application specific management and data related management. From the two key factors mentioned above, two important conclusions can be drawn about the relationship of domain scientists with these management challenges. One, their main focus is on handling the first dimension (Application Management) and thus they are comparatively less intended towards the management of the second dimension (Data Management). Two, they are experiencing comparatively more problems in managing the second dimension (Data Management). As a result, handling both the intricate dimensions by scientists, in an unstructured way, results in focus divergence.

Therefore, Independent and separate specification of both application and data management dimensions would help them reduce the workflow complexities. Thus, a mechanism that implements conceptual separation between both management dimensions is direly needed.

II. Related Works

A scientific Workflow Management System (SWfMS) is the framework that offers the means to completely define, manage, monitor, and execute the scientific experiments in terms of scientific workflows. The design of a generic architecture at an appropriate level of abstraction that properly and transparently addresses the essential requirements for SWMSs is critical and challenging.

The formal concept of workflow has existed in the business world for a long time. The Workflow Management Coalition (WfMC) [1] has proposed a reference architecture for business workflows. Since then, the reference architecture and its variants [2] have been widely adopted in development of business workflow management systems [3, 4, 5, 6]. However, in [7], authors have convincingly argued that these reference architectures are not suitable for scientific workflow management systems since scientific workflows have different goals.

During the past years, several scientific workflow management frameworks have been emerged [8, 9], which offer fairly much experiences for future research and development. In this section we will report some popular systems and also provide a comparative discussion.

Kepler [10, 11] is one of the popular open source scientific workflow systems with contributors from a range of application-oriented research projects such as Ecology [12], Biology [13], and Geology [14]. Kepler is built on Ptolemy II, a PSE (Problem Solving Environment) from electrical engineering and thus inherits actor-oriented feature from it. Kepler is dataflow-centric and uses proprietary modeling language so called MoML [15] for workflow specification.

Taverna [16, 17] is also an open source scientific workflow management system like Kepler; it is a part of myGrid project which aims to employ Grid technology to develop high level middleware for supporting personalized in silico experiments [18] in biology. Taverna is implemented as a service-oriented architecture based on Web Service standard, thus the data channel between two services works on SOAP based XML messages.

Triana [19, 20] is an open source workflow based graphical problem solving environment PSE, aiming at defining, analyzing, managing, executing and monitoring the workflows that handles a range of distributed elements such as grid jobs, web services and P2P communication. Although Triana was developed for data-analysis scientists in GEO 600 project, it can be used in many different ways and a rich library of units currently exits covering a broad range of applications.

Pegasus [21] is a framework that can manage the execution of complex scientific workflows on distributed resources. Pegasus is the part of GriPhyN [22] project which aims at supporting large-scale data management in physics experiments such as astronomy, high energy physics, and navigation wave physics. Pegasus enables scientists to design workflows on application level without the need to worry about the actual execution environment. Thus abstract workflows designed by the domain scientists are independent of any resources they will be executed on. Pegasus basically provides the functionality to map the scientific workflows onto distributed resources at a Grid middleware.

ASKALON [23, 24] is a framework developed for Grid application development and computing environment. Its ultimate goal is to simplifying the development and optimization of scientific workflows that can harness the power of Grid computing. In ASKALON, workflows are defined using its property XML-based language known as Abstract Grid Workflow Language AGWL. Like Pegasus, the language enables users to define workflows on abstract level without involving into the middleware complexities and dynamic nature of the Grid. Workflows are composed by using atomic units of works called Activities interconnected through control-flow and data-flow dependencies. The language provides a rich set of constructs to express sequence, parallelism, choices, and iteration workflow structures.

SODIUM [25] (Service Oriented Development In a Unified FraMework) is a platform which provides a set of languages, tools, and corresponding middleware, for modeling and executing scientific workflows composed of heterogeneous services. The system is implemented as Service-oriented architecture and the main objective is to provide seamless access to different types of services such as Web services, Grid Services and P2P services. The overall functionality is achieved by three phases. First, user need to model requirements for services which will satisfy specific workflow task.

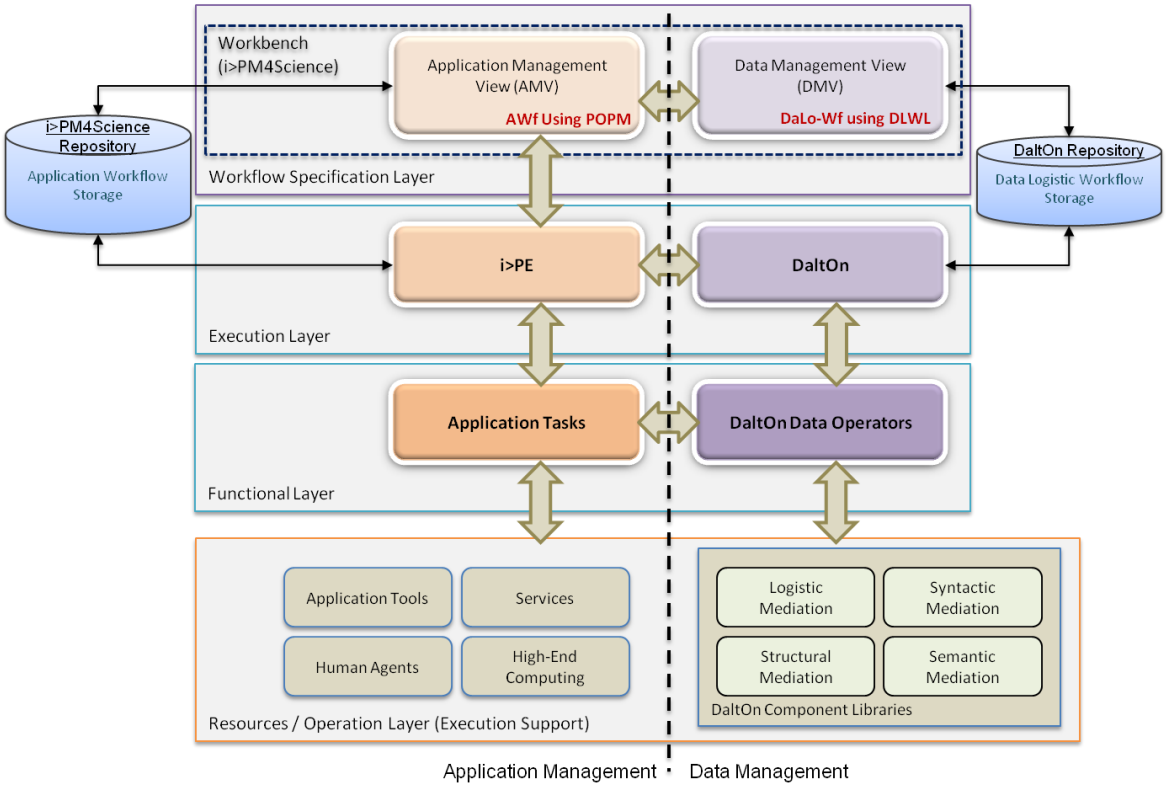


Fig. 1. Architecture of scienceFLOW (a Scientific Workflow Framework)

e-Bioflow [Wass08] is workflow design system which, considering the fact that workflow designers from different domains prefer different perspectives, enables users to model a workflow from three different perspectives: the control flow perspective, the data flow perspective, and the resource perspective. e-Bioflow is inspired by the context of scientific collaborative environment such as e-BioLab and relies on an existing system to have the workflow enacted; workflow models developed by e-Bioflow are enacted by the open-source workflow system Yawl [Van04].

These systems are well developed and are powerful in offering very rich libraries of pre-developed computational components, in executing workflows through the distributed and high computing environments such as Grids and P2P, in managing provenance information, in managing data on grid based technologies, and also in providing many novel and innovative features. However, an architectural reference that can provide a high level management of sub-systems and their interactions in a scientific workflow framework is still an open issue. The current systems have either not an explicit architectural design or the architecture is proprietary and restricted greatly by the legacy system that the frameworks are built upon [7]. For example, Kepler is built on the Ptolemy II, and hence, each new requirement that is needed to be incorporated by the framework is based on the extensions to the architecture of the underlying system, i.e. Ptolemy II. Pegasus, on the other hand, is built upon Condor and Dagman by adding another workflow mapper on the top of these two systems.

III. Scientific Workflow Framework – Science*FLOW*

In order to define workflow specifications of both application and data management dimensions in an independent way, the paper presents an integrated framework, called science*FLOW* [Figure 1] that implements both dimensions in a transparent manner. The framework is composed of four layers having various operational sub-systems at each layer.

A. Layers and Sub-systems

Figure 1 depicts the architectural view of our framework that integrates a number of operational sub-systems. In the following, we will provide a detailed discussion of each operational sub-system on each layer.

Workflow Specification Layer: On this layer, in order to define a scientific workflow we have built two separately graphical sub-systems '**Application Management View AMV**' and '**Data Management View DMV**' for specifying two categories of workflows 'Application Workflow *AWf*' and 'DataLogisitc Workflow *DaLo-Wf*' respectively. The '**Workbench',** known as *i>PM4Science*, of the framework integrates both sub-systems and thus provides two different graphical tabs (Views) [Figure 3] in order to represent them disjointedly. In this way, the workbench promotes two community groups to work together but on their corresponding tabs, i.e. scientists on '*AMV'* and data experts on '*DMV'*. In the first tab '*AMV*' [Figure 3(a)], '*AWf*' is defined using *POPM* notion [29] that allows scientists to express scientific operations on very abstract level without involvement into application and data technicalities.

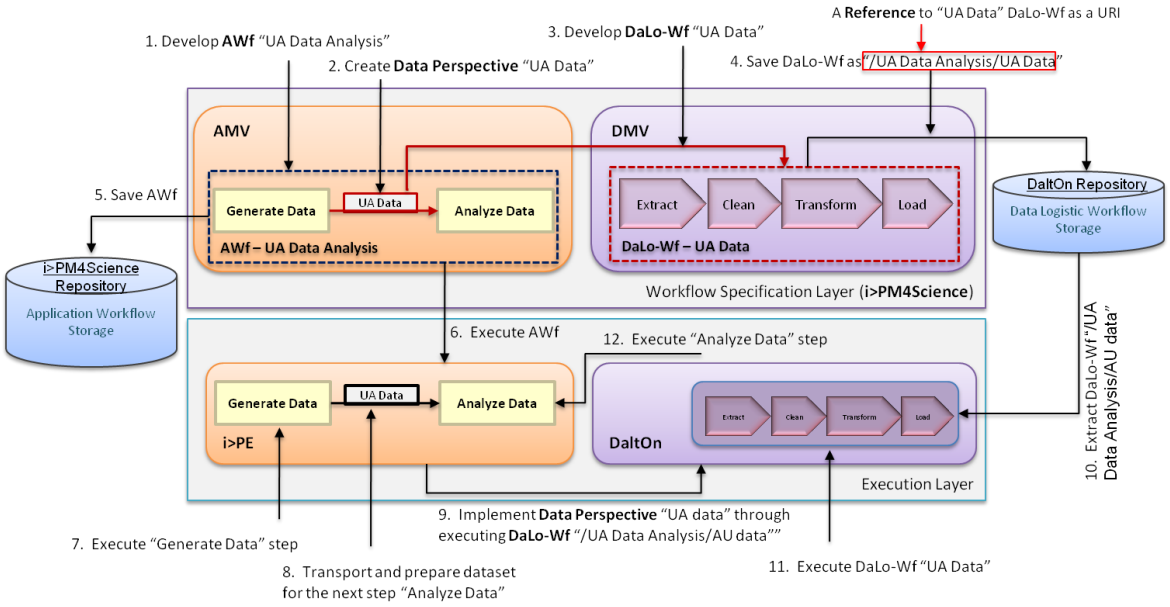


Fig. 2. A Simple Example – showing the course of actions and the flow of Information in the framework

In the second tab '*DMV*' [Figure 3(b)], for each of '*Data Perspective'* specified in '*AWf*' a '*DaLo-Wf*' is defined using a specialized (data centric) language.

Figure 2 depicts a simple example for developing and executing a scientific workflow through the framework, demonstrating the course of actions and the flow of information throughout the sub-systems in the framework. The course of actions consists of 12 stages and the '*Workflow Specification Layer*' covers 1 to 6 stages. As an example, a scientist develops a scientific experiment for analyzing weather data (UA dataset) stemming from a sensor device (Ultrasonic Anemometer). At first stage, the scientist develops an '*AWf*' (named "UA Data Analysis") using '*AMV',* including only two experiment related steps '*Generate Data*' and '*Analyze Data*'. In the second stage, (s)he creates a '*Data Perspective*' (named "UA Data"). At the third stage, a '*DaLo-Wf*' (named "UA Data") is developed - against the created '*Data Perspective'* - using '*DMV*', including data related steps '*Extract*', '*Clean*', '*Transform'* and '*Load'*. After successful specification of the '*DaLo-Wf',* at the fourth stage it is stored into '*DaltOn Repository'* at the location "/UA Data Analysis/UA Data". In this way '*Data Perspectives*' defined into '*AWf*' do not contain the concrete specification of respective '*DaLo-Wfs*'; rather holds a **Reference** of the location of such workflows. By default the name of '*DaLo-Wf*' in the '*DaltOn Repository*' is the same as that of '*Data Perspective*' in the '*AWf*' with the contextual path defined in the '*AWf*', nevertheless users can change it. Moreover, the '*Data Perspective'* can refer already created '*DaLo-Wf'* by just annotating it with the reference of the location of a particular workflow. In this way the reusability of '*DaLo-Wfs'* in multiple application workflows can easily be achieved. Basically, the reference of the location is a URI and thus can be qualified with the address of the system where '*DaltOn Repository'* resides, for instance "http://132.180.195.110/UA Data Analysis/UA Data"; by default the repository is expected to be at the local system. At the fifth stage the complete specification '*AWf*' is stored into the '*i>PM4Science Repository*' for future use. At the sixth stage the complete '*AWf*' specification is passed to the execution environment (*i<PE*) for executing it.

The screenshots of our implemented workbench are shown is Figure 3 where (a) represents '*AMV*' and (b) depicts '*DMV*'. Since all the *DaLo-Wfs* are stored into '*DaltOn Repository'*, '*DMV'* also provides the ability to search - from the repository - already defined '*DaLo-Wfs'*. '*DMV*' fundamentally consists of three elements, i.e. design panel (right) for defining '*DaLo-Wf*' using *DLWL* notion, *Data Perspective* explorer panel (left) for browsing through '*DaLo-Wfs'*, and search panel (upper left) for searching previously defined '*DaLo-Wfs'*.

Execution Layer: At this layer two dedicated sub-systems are employed, i.e. '**i>PE**' [30] and '**DaltOn**' [31] for executing both categories of workflows separately and independently. The sub-system '*i>PE*' (**i**ntegrated **P**rocess **E**xecutor) is a process-centric workflow engine founded on *POPM* paradigm, that makes available the environment for executing '*AWf*' specified via *POPM* notion. During the course of '*AWf*' execution, whenever the engine comes across '*Data Perspective'* between two work steps it invokes the contiguous sub-system, i.e. '*DaltOn'*, in order to implement data perspective under the focus. The sub-system '*DaltOn*' is a data processing system which is specifically designed to provide the data management mediation to the scientific workflows by implementing data management part (*Data Perspective*).

The Figure 2 (lower layer) demonstrates the course of actions and the information flow through the execution sub-systems employed at this layer. The course of actions at this layer encircles 6 stages (from 7 to 12). At the seventh stage, the execution engine '*i>PE*' starts executing the '*AWf*' through the invocation of the first work step '*Generate UA Data*' that aims at extracting the dataset from sensor device. At the eighth stage, the engine flags the next step '*Analyze Data*' as an "executable" and identifies the data transportation and preparation task for the step. Then at the ninth stage, in order to

implement the '*Data Perspective'* (i.e. "UA Data") the engine invokes and requests the '*DaltOn'* to execute the respective '*DaLo-Wf*', by passing the location reference of the workflow (i.e. "/UA Data Analysis/UA Data"). After getting the execution request from '*i>PE*', the '*DaltOn'* extracts the corresponding '*DaLo-Wf*' (graph of data processing tasks) from '*DaltOn Repository*' at the tenth stage and implements the '*Data Perspective'* (i.e. "UA Data") by executing the respective '*DaLo-Wf'* at the eleventh stage . In this way '*DaltOn Repository*' plays an interface role between '*DMV*' and '*DaltOn*' sub-systems. Finally at the twelfth stage, '*i>PE*' executes the last step (i.e "Analyze Data") in the example '*AWf*'.

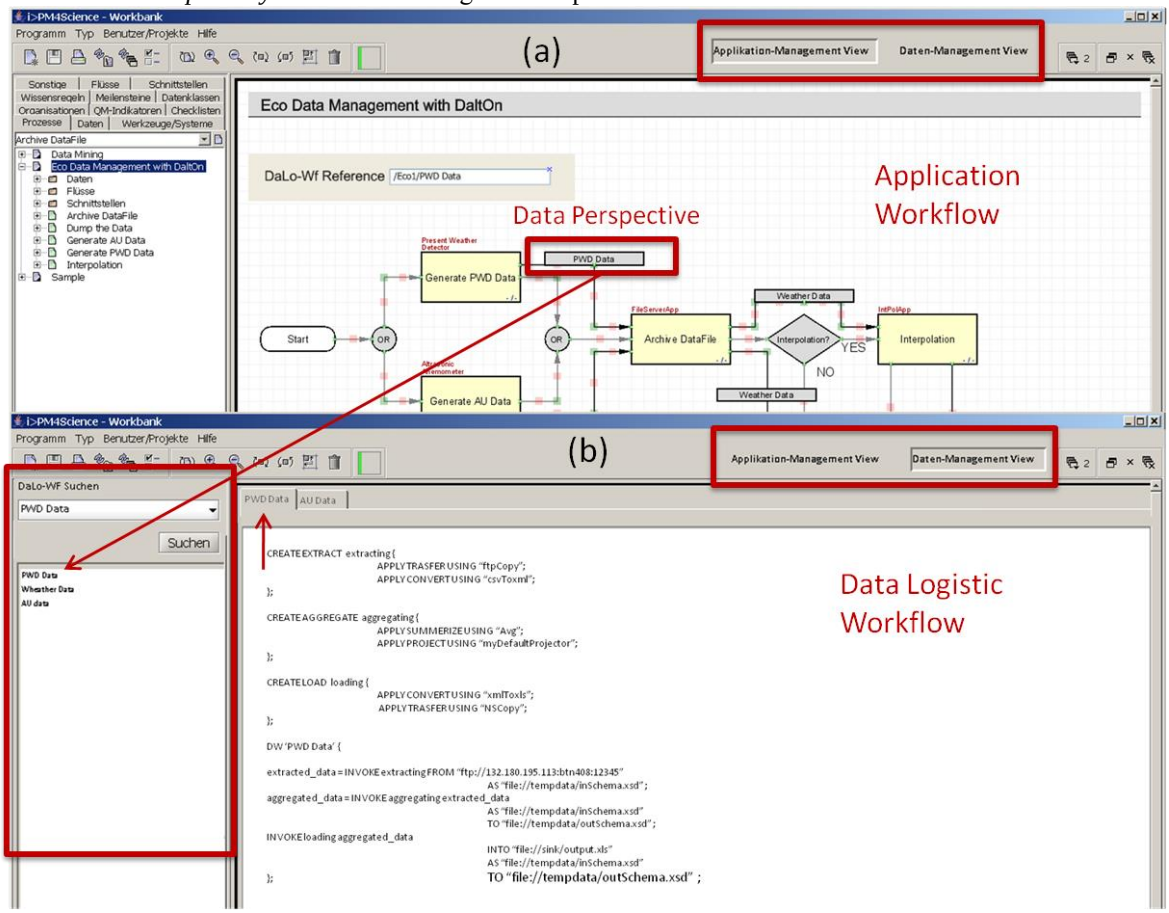


Fig. 3. i>PM4Science Workbench – (a) Application Management View - AMV, (b) Data Management View - DMV

Functional Layer: On this layer, in order to provide a generic execution support to both sub-systems on '*Execution*' layer we built up two sub-systems, i.e. '**Application Tasks**' and '**DaltOn Data Operators**'. Basically, both the sub-systems constitute the libraries for different kind of logical operations (tasks) with the purpose of supplying execution entities to the execution engines.

'*Application Tasks*' sub-system aims at providing the library of commonly used experiment related tasks that play the role of building blocks for '*AWf*' development. The sub-system also supports some features that are required for managing and maintaining a component library, for instance adding, removing and editing the experimental tasks. The tasks provided by such sub-system are completely abstract entities, thus the physical implementation of each of these tasks is realized by the underlying concrete functions or applications. Therefore, a single task can be utilized with multiple implementations in a generic way.

'*DaltOn Data Operators*' aims at constituting the library of the most common data operations that play the role of basic building blocks for '*DaLo-Wf*' development. Like above sub-system, it also supports some basic features that are needed for managing and maintaining a component library. Data operators offered by this sub-system are logical entities, thus the physical implementation of each of these operators is realized by the underlying concrete functions. Therefore, a single operator can be utilized with multiple implementations in a generic way. For instance, the operator '*Convert*' reflects an abstract operation that can be utilized in a number of implementations such as '*csv2xml*', '*ua2xml*', and '*xml2xls*' - by just providing a specific low level function.

Resource / Operational Layer: Fundamentally, this layer aims at supplying physical resources to the upper layer in order to implement the logical and abstract experimental tasks as well as data operators. In order to support experimental tasks we maintain the libraries of application / software tools, services such as Web or Grid services, human agents. In order

to support the data operators we implemented comprehensive and classified component libraries of diverse functions.

IV. CONCLUSION

The paper presented a generic and scalable scientific workflow framework *'scienceFLOW'*, whose originality is a clear isolation of two management concerns, i.e. application and data management. The design solution of the framework was motivated by two factors, i.e. the identified requirements and the method for e-Science. The basic 'requirement' (i.e. application specific and data related issues should be handled entirely in a separate manner) is nicely fostered in the framework since the whole framework is divided into two segments, i.e. the application management and the data management. In order to manage the issues occurring in both segments dedicated sub-systems are employed not only at the design level but also at the execution level.

REFERENCES

[1] Workflow Management Coalition: "The Workflow Management Coalition (Homepage) ". http://www.wfmc.org, [Oct 2016]

[2] Grefen, P. and de Vries, R.: A Reference Architecture for Workflow Management Systems," Data Knowledge Eng., vol. 27, no. 1, pp. 31-57, 1998

[3] YAWL: Yet Another Workflow Language: "YAWL Foundation (Homepage)". http://www.yawlfoundation.org/ [Oct 2016]

[4] van der Aalst, W., Aldred, L., Dumas, M. and ter Hofstede, A.: Design and Implementation of the YAWL System, Proc. Center for Advancement of Informal Science Education Conf. (CAiSE '04),pp. 142-159, 2004

[5] Wil M. P. van der Aalst: Business Process Management: A Comprehensive Survey, Hindawi Publishing Corporation, ISRN Software Engineering, Volume 2013

[6] Ming Gao, Lei Yang, Chunhua Zhang, Wen Guan and Ailing Li: Towards Unified Business Process Modeling and Verification for Role-based Resource-oriented Service Composition, International Journal of Hybrid Information Technology Vol.9, No.3 (2016), pp. 145-158, 2016

[7] Lin, C.; Lu, S.; Chebotko, A.; Pai, D.; Lai, Z.; Fotouhi, F. and Hua, J.: A Reference Architecture for Scientific Workflow Management Systems and the VIEW SOA Solution. IEEE Transactions on Services Computing, Vol. 2(No. 1): p. 79-92, 2009

[8] AL Lemos, F Daniel, B Benatallah: Web Service Composition: A Survey of Techniques and Tools, ACM Computing Surveys, Vol. 48, No. 3, Article 33, December 2015

[9] Scientific Workflow Survey (home page), http://www.extreme.indiana.edu/swf-survey, [Oct 2016]

[10] Daniel Crawl, Alok Singh, and Ilkay Altintas: Kepler WebView: A Lightweight, Portable Framework for Constructing Real-time Web Interfaces of Scientific Workflows, Procedia Computer Science, Volume 80, Pages 673–679, 2016

[11] Ludäscher, B.; Altintas, I.; Berkley, C.; Higgins, D.; Jaeger-Frank, E.; Jones, M.; Lee, E.; Tao, J.; Zhao, Y.: Scientific Workflow Management and the Kepler System. In Concurrency and Computation: Practice & Experience, Vol. 18 No. 10, Citeseer, 2006

[12] SEEK: Science Environment for Ecological Knowledge (home page), www.seek.ecoinformatics.org, [Oct 2016]

[13] Clotho: Design environment for synthetic biological systems (home page), http://www.clothocad.org/, [Oct 2016]

[14] GEON: The NSF Information Technology Research (ITR) program, http://www.geongrid.org, [Oct 2016]

[15] Lee, E. A. and Neuendorffer, S.: MoML — A Modeling Markup Language in XML — Version 0.4, Technical Memorandum UCB/ERL M00/12, University of California, Berkeley, CA 94720, March 14, 2000

[16] Katherine Wolstencroft1, Robert Haines, Donal Fellows: The Taverna workflow suite: designing and executing workflows of Web Services on the desktop, web or in the cloud, Nucleic Acids Research Advance Access, Oxford University Press, May 2013

[17] Taverna: An open source scientific workflow system (home page), http://www.taverna.org.uk, [Oct 2016]

[18] Oinn, T.; Addis, M.; Ferris, J.; Marvin, D.; Greenwood, M.; Goble, C.; Wipat, A.; Li, P.; Carver, T.: Delivering web service coordination capability to users, Proceedings of the 13th international World Wide Web conference on Alternate track papers & posters, p. 438-439, ACM New York, NY, USA, 2004

[19] Taylor, I. J.; Deelman, E.; Gannon, D. B.; Shields, M.: Workflows for e-Science: Scientific Workflows for Grids, Springer-Verlag London Limited, 2007

[20] Triana - Open Source Problem Solving Software (home page), www.trianacode.org [Oct 2016]

[21] Ewa Deelmana, Karan Vahia, Gideon Juve: Pegasus, a Workflow Management System for Science Automation, Future Generation of Computer Systems, Elsevier, Volume 46, Pages 17–35, May 2015

[22] Grid Physics Network (home page): http://www.usatlas.bnl.gov/computing/grid/griphyn, [Oct 2016]

[23] Fahringer, T: ASKALON Grid Environment, User Guide Version 1.2, Distributed and Parallel Systems Group, Institute of Computer Science, University of Innsbruck, September 2015

[24] Fahringer, T.; Prodan, R.; Duan, R.; Hofer, J.; Nadeem, F.; Nerieri, F.; Podlipnig, S.; Qin, J.; Siddiqui, M.; Truong, H. L.: Askalon: A development and grid computing environment for scientific workflows, Workflows for eScience, Scientific Workflows for Grids, Springer, 2007

[25] Tsalgatidou, A.; Athanasopoulos, G.; Pantazoglou, M.; Pautasso, C.; Heinis, T.; Gronmo, R.; Hoff, H.; Berre, A. J.; Glittum, M.; Topouzidou, S.: Developing scientific workflows from heterogeneous services, ACM Sigmod Record, Vol 35, No. 2, ACM, 2006

[26] Wassink, I.; Rauwerda, H.; van der Vet, P. E.; Breit, T.; Nijholt, A.; Elloumi, M.; Küng, J.; Linial, M.; Murphy, R. F.; Schneider, K.: e-BioFlow: Different perspectives on scientific workflows, 2nd Internationall Conference on Bioinformatics Research and Development (BIRD), Springer, 2008

[27] The First International Workshop on Data and Process Provenance WDPP (homepage http://www.itee.uq.edu.au/~dasfaa/workshop/wdpp/WDPP09.htm), 2009

[28] Greenwood, M.; Goble, C.; Stevens, R.; Zhao, J.; Addis, M.; Marvin, D.; Moreau, L.; Oinn, T.: Provenance of e-Science experiments - experience from bioinformatics, In Cox, S., ed.: Proceedings of UK e-Science All Hands Meeting 2003, Nottingham, UK. 2003

[29] Jablonski, S., Bussler, C.: Workflow management. Modeling concepts, architecture and implementation, Thomson, London, UK, 1996

[30] Jablonski, S.; Faerber, M.; Götz, M.; Volz, B.; Dornstauder, S.; Müller, S.: Configurable Execution Environments for Medical Processes, 4th Int'l Conf. on BPM, BPM Demo Session 2006, Vienna, 2006

[31] Jablonski, S.; Curé, O.; Rehman, M. A.; Volz, B.: Architecture of the DaltOn Data Integration System for Scientific Applications, Proceedings of the 2008 Eighth IEEE International Symposium on Cluster Computing and the Grid, IEEE Computer Society, 2008.

Efficient Relay Selection Scheme based on Fuzzy Logic for Cooperative Communication

Shakeel Ahmad Waqas
Military College of Signals
National University of Sciences and Technology (NUST)
Rawalpindi/Islamabad, Pakistan

Nasir Khan
Department of telecommunication Engineering
University of Engineering and Technology (UET)
Mardan/Peshawar, Pakistan

Imran Touqir
Military College of Signals
National University of Sciences and Technology (NUST)
Rawalpindi/Islamabad, Pakistan

Imran Rashid
Military College of Signals
National University of Sciences and Technology (NUST)
Rawalpindi/Islamabad, Pakistan

***Abstract*—The performance of cooperative network can be increased by using relay selection technique. Therefore, interest in relay selection is sloping upward. We proposed two new relay selection schemes based on fuzzy logic for dual hop cooperative communication. These relay selection schemes require SNR (signal to noise ratio), cooperative gain and channel gain as input fuzzy parameters for selection of best relay. The performance of first proposed relay selection scheme is evaluated in term of BER (bit error rate) in Nakagami, Rician and Rayleigh fading channels. In second proposed relay selection scheme, threshold is used with the objective to minimize the power consumption and channel estimation load. Its performance is analyzed in term of BER, number of active relays and load of number of channel estimations.**

Keywords—Cooperative Networks; Relay selection schemes; Amplify and forward; Fuzzy logic; Nakagami Fading channel; Rician Fading Channel; Rayleigh Fading Channel

I. INTRODUCTION

In future wireless communication networks, multipath fading is the key problem to achieve high data rate. Time, frequency and spatial diversity techniques are considered to alleviate the multipath fading. Taking specifically spatial diversity, which normally improves the system by introducing independent path communication. Primitively, the concept was achieved by introducing multiple antennas (MIMO) which improves the performance manifolds. Due to size, cost along with some hardware limitations, the system practical implementation becomes another problem itself. For overcoming these concerns, cooperative communication was introduced as a virtual MIMO environment [1]. For bringing independent path communication into being for practical implementation and exploiting the broadcast nature of the wireless communication destination relays are introduced between source and destination, which in literal sense are not more than forwarders of the source signal based on some designed protocol [2].

Relaying protocols are followed by the relay for forwarding of the signal to destination which includes namely "Amplify and Forward" (AF), "Decode and Forward" (DF), "Estimate and Forward", "Compressed and Forward" etc. [3]. For minimizing complexity of the system, AF is the best technique which forward the amplified received signal of the source to destination with the demerit of noise also being get amplified along with the signal [4] [2] [5] [6] [7, 8]. For better performance DF is normally taken into account which decodes the signal at the relay and then encode it back before forwarding [2, 4] [9] [10] [11]. In practical sense, more than one relays are present and use of all relays will leads to interference with the sources, high power consumption and consuming high bandwidth. For covering up this problem a relay with better specifications and according to the requirements of the application is chosen which forwards the signal from source to destination.

Relay selection is studied extensively these days by researchers and magnificent work can be found in [12-28] .After analysis the literature, we can in general classify the work done on relay selection into five categories based on the kind of selection technique used:

a) Geographical information based relay selection

Geographical information based relay selection with the aim to minimize symbol error probability was presented by Wang et al. which used the distance from source to relay and from rely to destination as a criterion for best relay selection. But because of channel fading and shadowing effect the proposed algorithm is not applicable for practical scenario[12]. Three-time slot TDMA based transmission protocol has been investigated and analyzed by U. R. Tanoli et al. with the better performance than old protocols using the location information. The deficiency of the protocol is an extra time slot for transmission thus decreasing the code rate and also impractical because of location based information [13]. Analysis of AF and DF relaying protocols for inter-relay communication with the proof that AF can perform better than DF in term of BER is presented in [14].

b) Energy efficiency based relay selection

The expression of total energy is obtained and used for relay selection in general scenario with characterization of the structure for optimal transmission [15]. Power aware relay

selection is proposed by Yan Chen et al. with the aim of minimizing the overall network life-time. Furthermore, optimal power allocation and three power-aware selection criteria are being analyzed [16]. Relay selection on the basis of channel state information with the contribution of power allocation algorithm, for the purpose of presenting an energy-saving relay selection strategy. The solution presented is well analyzed but location based information is also used. However computational complexity and requirement of information at the source is not fully addressed [17].

c) Outage probability based relay selection

Optimal-outage relay selection scheme is presented by Li Sun et al. with utilization of feedback from the receiver for the decision of whether rely cooperation is required or not. Adaptive DF and AF is used in this paper as forwarding scheme [18]. However, outage priority based fairness is proposed by Li Yubu et al. with the aim to improve relay selection fairly without performance disturbance and with improvement in network lifetime [19]. The effect of correlated log-normal shadowing is analyzed based on outage probability using opportunistic DF and showing significant impact on outage performance [20].

d) Interference aware relay selection

Interference is pretty common to exist in wireless communication and have an unavoidable effect over the performance of the communication in case of multiple transmission pairs. Relay Interference effect in cooperative networks is analyzed by Y. Zhu and H. Zheng between interference management and cooperative relay strategy and two spectrum selection techniques are presented with different tradeoffs. However, the paper does not deal the problem of relay selection at all [21]. Furthermore, interference based relay selection is devised with the purpose of maximizing the mutual information on cooperative networks with the interference limited destination [22]. Interference aware relay selection is proposed with the use of distributed interference aware relay selection algorithm by C Shi et al in [23] to select best relay by using inter node interference and channel statistics.

e) Channel state information based relay selection

Opportunistic relaying was presented in [24] as the best scheme for relay selection in which the source have exact information of source to relay and then relay to destination, only then source was able to select the best relay out of all the relays. The system delay during selection and not exact surety of same channel during data transmission as was during estimation were the major issues not addressed in the paper. The authors in [25] have utilized the outdated channel state information for selection of best relay and have adopted maximum a posteriori for the prediction of actual SNR during the transmission and have utilized it for single relay selection as a strategy. An algorithm is proposed in [26] in which instead of using global knowledge of all the paths local channel states available at the relay is utilized for relay to select and mark itself as the best relay thus implementing distributed relay selection. SNR is calculated using outdated method of SNR calculation and then according to length of data transmission and channel state information available best relay is selected thus reducing the repeated relay selection technique [27]. Furthermore, Glauber Brante et. al. reported a fuzzy logic based relay selection algorithm for wireless cooperative sensor network. CSI (channel state information) and residual energy are the input fuzzy parameter of fuzzy controller [28].

To the best of our knowledge, the discussed relay selection techniques are causing enormous control message overhead because of repeatedly relay selection for cooperation, when successive data transmission is carried out especially in case of audio or video session occurring between source and destination, thus causing degradation in network performance. To reduce the frequent selection of relays we propose two algorithms based on local channel state (CSI) information available with the use of fuzzy logic algorithm for computational complexity and control message overhead reduction while promising for selection of best relay and guaranteeing betterment in symbol error rate. Three fuzzy parameters are used in the defined algorithms namely signal to noise ratio (SNR), cooperative gain and channel gain. Cooperative gain is defined here as the ratio of direct transmission bit error rate (BER) to cooperative transmission BER. The performance of proposed relay selection scheme is analyzed and compared on dual hop cooperative network over three fading channels namely Rayleigh, Rician and Nakagami. For reduction of control message overhead and power consumption, another modified algorithm is proposed with the use of threshold. Moreover, number of active relays and number of channel estimation depend on threshold chosen. The tradeoff curves of threshold, BER, number of active relays and channel estimation are presented in this paper. AF is used as relaying protocol and MRC (maximal ratio combining) is selected as combining technique at destination. The performance of both relay selection schemes are evaluated through Monte Carlo simulation. Comparison with previous relay selection schemes is also carried out in this research.

Rest of the paper is organized in the following pattern. Section 2 defines the system model proposed in this research. Whereas relay selection criterion and selection algorithms are discussed in section 3. Section 4 comprises of the simulation model and simulation parameters while simulation results are discussed in section 5. Paper is concluded and future work is presented in section 6.

II. System Model

Considering the cooperative communication, our proposed system model investigated in this paper consist of a source (S) transmitting its signal to destination (D) through the cooperation of N relays R_i wheras $(i = 1,2,3 \dots N)$ as shown in Fig. 1. with the consideration that each terminal is mounted with single antenna. An assumption is taken that all the relays are operating in half duplex mode and let h_{s-d}, h_{s-ri} and h_{ri-d} are the channel coefficients from source to destination, source to i^{th} relay and form i^{th} relay to destination respectively. It is also assumed that zero mean white Gaussian noise is there at each terminal. As The network model is dual hop with the consideration that no inter relay communication is occurring and the source knows the channel statistics of both the relays using the loop feedback. Considering E_s as transmission power of the source while E_{r_i} as the transmission power of the i^{th} relay. Now let X be the signal transmitted by source and Y_D is

the signal received at destination then as discussed in [2] the signal received at destination can be represented as:

$$Y_D = h_{s-d} * \sqrt{E_s}\, X + n_d \quad (1)$$

Whereas n_d is the noise at destination. Now the signal received at i^{th} relay will be:

$$Y_{r_i} = h_{s-r_i} * \sqrt{E_s}\, X + n_{r_i} \quad (2)$$

Now as according to the considered relaying protocol AF, an amplified copy of the received signal amplified with a factor of β at i^{th} relay will be forwarded towards destination which can be represented as:

$$Y_{D-r_i} = h_{r_i-D} * \sqrt{E_{r_i}}\, \beta \left(h_{s-r_i} * \sqrt{E_s}\, X\right) + h_{r_i-D} * \sqrt{E_{r_i}}\, \beta\, n_{r_i} + n_D \quad (3)$$

Two time slots model is considered for avoidance of interference in which signal is broadcasted by the source at time slot 1 and received by both the destination and the i^{th} relay. While in 2nd time slot amplified copy of the received signal is transmitted by the relay. Multiple copies of source signal are received at destination and are combined together using MRC technique while ignoring any time delay occurred during transmission. For simplicity, we assume the path loss component equal to 1. Using proposed relay selection algorithm, best relay out of all will be selected for cooperation by the destination and the information will be sent to all the relays along with source using feedback transmission.

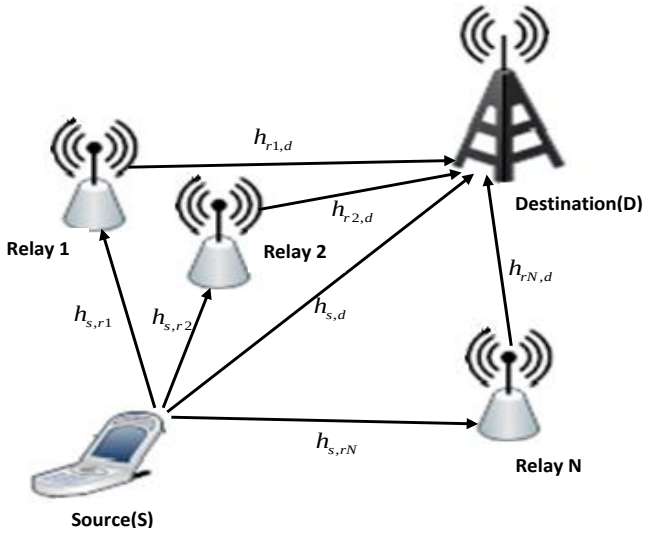


Fig. 1. System Model

III. Relay Selection Criterion and Algorithms

A cooperative communication model was discussed in the previous section. Now the scenario is that we will be having *N* number of relays. Out of all the available relay one relay which can outperform other on the basis of performance will be selected for cooperation by using the algorithms discussed here. But before discussing relay selection algorithms first we will discuss the following parameters which are used for relay selection on the basis of the defined algorithms:

B. SNR

The SNR of direct path i.e. transmission of signal from source to destination is denoted by γ_{S-D} and can be written as[29]:

$$\gamma_{S-D} = \frac{|h_{S-D}|^2}{N_D} \quad (4)$$

The SNR of 1st hop of i^{th} relay is denoted by γ_{S-r_i} and its equation is:

$$\gamma_{S-r_i} = \frac{|h_{S-r_i}|^2}{N_{r_i}} \quad (5)$$

The amplified signal is forwarded to destination in 2nd hop. The SNR of 2nd hop of i^{th}relay is represented by γ_{r_i-D} and is given as:

$$\gamma_{r_i-d} = \frac{\beta |h_{S-r_i}|^2 |h_{r_i-D}|^2}{(\beta^2 |h_{S-r_i}|^2 + 1) N_{r_i}} \quad (6)$$

For the relay selection, the minimum SNR of two hop is used. The SNR value of i^{th} relay used for relay selection is given below:

$$\gamma_i = \min(\gamma_{S-r_i}, \gamma_{r_i-d}) \quad (7)$$

C. Cooperative Gain

Cooperative gain is defined as the ratio of BER of direct transmission to BER of cooperative transmission. The BER of i^{th} relay used for relay selection can be written as:

$$CG_i = \frac{BER_{direct}}{BER_{Cooperative}} \quad (8)$$

D. Channel Gain

For the proposed relay selection scheme, the minimum channel gain of two hop is considered. For i^{th} relay, the channel gain used for relay selection is:

$$h_i = \min(h_{S-r_i}, h_{r_i-d}) \quad (9)$$

The relay selection scheme proposed in this work is making use of fuzzy logic, which consists of input fuzzy parameters, fuzzification, fuzzy inference system, fuzzy rules and defuzzification. SNR, cooperative gain and channel gain are the three input fuzzy parameters. In the process of fuzzification, a static value is assigned to variables γ_i, CG_i and h_i defined by the input membership functions. SNR is having five membership functions $(V.Low, Low, Medium, High, V.High)$, cooperative gain is having three membership functions $(Poor, Normal, Good)$, while channel gain is also having three input membership functions $(Bad, Moderate, Best)$ and by calculation total 45 rules are defined for it. The degree of relevance $f(\gamma_i, CG_i, h_i)$ is calculated considering the strength of each rule and the output membership functions (Not Selected, Considered, Selected) which are shown in Fig. 3. The higher $f(\gamma_i, CG_i, h_i)$, higher will be the "quality" of the selected relay. Both the algorithms are shown in Fig. 2. Along with block diagram of relay selection scheme.

Algorithm I:
Step 1: Initiate $i = 0$
Step 2: Increment i by 1.

Step 3: Calculate the fuzzy parameters for i^{th} relay.
Step 4: Find out the degree of relevance
Step 5: If $i = N$, then jump to step 6, else jump to step 2.
Step 6: Compare the degree of relevance of each relay and select the relay having highest degree of relevance.
End

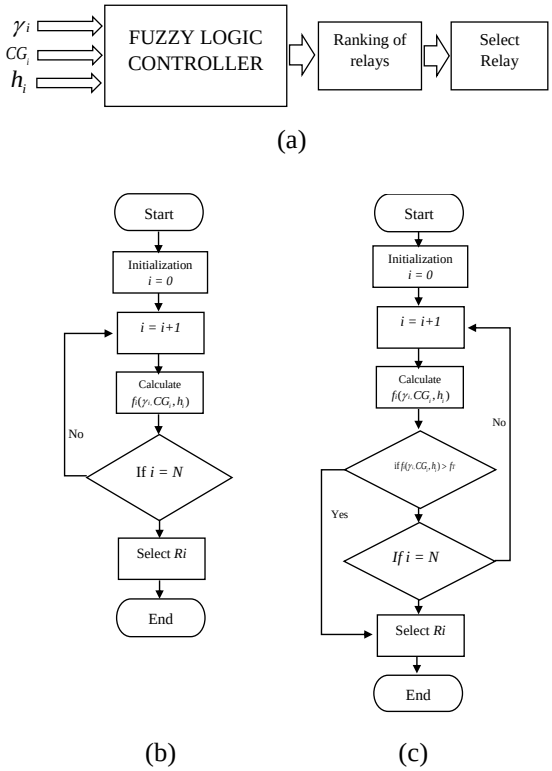


Fig. 2. (a) Block diagram of relay selection scheme (b) *Algorithm-1*(c) *Algorithm-2*

In this algorithm, 2N channels estimation must be conducted for selecting best relay and all the relays need to be in ON state. In order to reduce the load of channel estimation and power consumption, another relay selection algorithm is presented, in which use of predetermined threshold f_r is carried out. This scheme has four advantages over the previous. Channel estimation load i.e. number of relays whose path coefficients are taken into account, is reduced, no need to turn ON all the relays, time for relay selection is reduced and calculations for all the relays is reduced. Steps of this algorithm are given below:

Algorithm II:
Step 1: Initiate $i = 0$
Step 2: Increment i by 1
Step 3: Calculate the fuzzy parameters for i^{th} relay
Step 4: Find out the degree of relevance $f(\gamma_i, CG_i, h_i)$
Step 5: If $f(\gamma_i, CG_i, h_i) > f_r$, stop the calculation and go to step 8, else continue
Step 6: Else-if $i = N$, then continue, otherwise jump to step 2
Step 7: Compare the degree of relevance of each relay and select the relay having highest degree of relevance
Step 8: R_i is selected as best relay
End

Mamdani fuzzy inference system (FIS) [30] is used in both the algorithms for fuzzification. All the defined rules have equal weightage in calculating the degree of relevance and for defuzzification centroid approach is used.

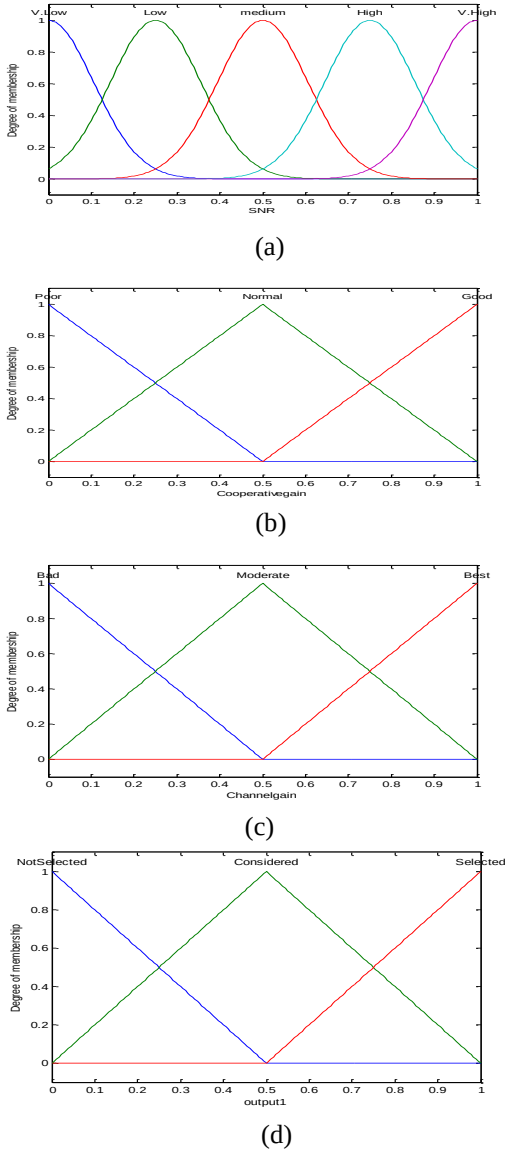


Fig. 3. Membership function (a) SNR (b) CGi (c) hi (d) degree of relevance

IV. Simulation Model and Parameters

In simulation model as shown in Fig. 4. we can observe that total of 5 relays are considered in-between source and destination. The model is showing the link from source to destination, source to relay and relay to destination. Each link is shown with its channel coefficient and the additive white Gaussian noise is also shown. Monte Carlo simulation are carried out for both the algorithms using the same simulation model. Amplify and forward as relaying protocol while B-PSK as a modulation scheme. During simulation transmission power was kept constant of 1watt for every node while the

amplification factor is kept fixed. Additive white Gaussian noise is used at each node and maximal ratio combining is used as receiver diversity whereas 10^6 bits are transmitted from source to destination for every simulation carried out. Both the algorithms are analyzed for three channel models namely Rayleigh, Nakagami and Rician. For Nakagami fading channel m is assigned value of 3 while k was assigned value of 5 for Rician fading channel. All these parameters are summarized in Tab. 1.

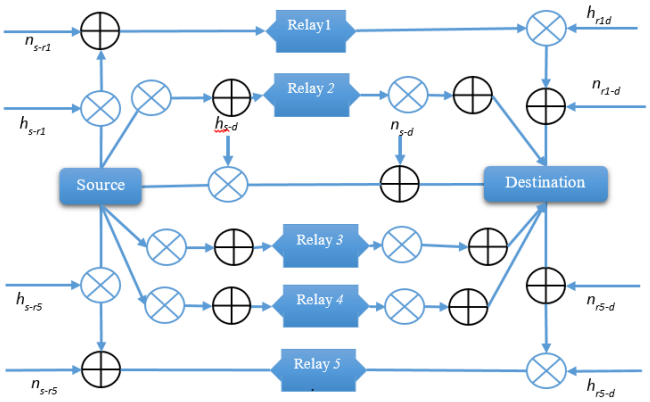


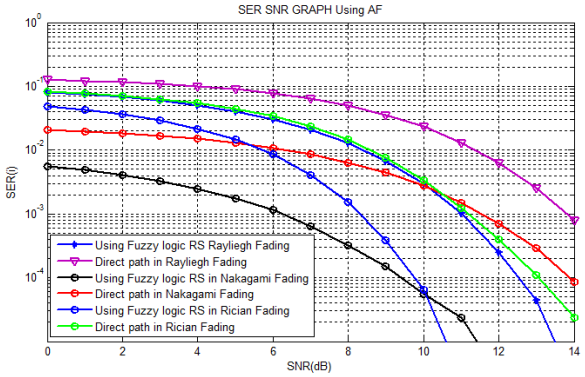
Fig. 4. Simulation Model

TABLE I. SIMULATION PARAMETERS

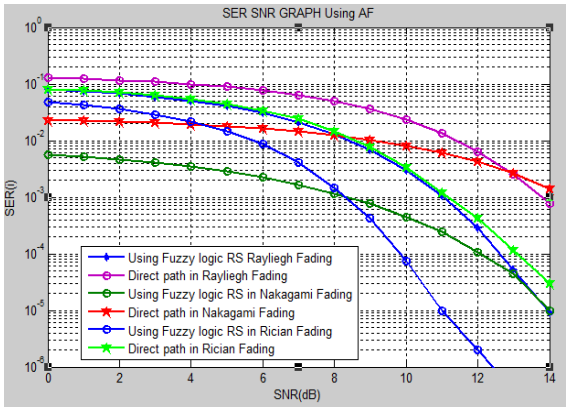
Parameters	Values
Source Transmit Power	1w
Relay Transmit Power	1w
Amplification Factor	Fixed
Modulation Scheme	B-PSK
Receiver Diversity	MRC
Number of Relays	5
Relaying Protocol	Amplify-and-Forward
Number of Bits	10^6
Receiver Noise	AWGN
Channel Model	Rayleigh, Nakagami, Rician

V. SIMULATION RESULTS AND DISCUSSION

Fig. 5.(a) shows the SER analysis of algorithm-1 over rayleigh, rician and nakagami fading channels. At low SNR values, better performance is observed in nakagami fading channel and at high SNR values, better performance is observed in rician fading channel. The reason behind this is the fact that is the line of sight component always have better performance at higher SNR regions. While Rayleigh which is more practical model, is having comparable results when cooperation is taken into account and compared with direct path of Rician. The SER analysis of *algorithm-II*, considering the threshold (f_r) equal to 0.7 is shown in Fig. 5.(b) is following exactly the same fashion as was observed in algorithm-1, reason for which is no change in channel models only reduction of computations and time of relay selection. The proposed relay selection, selected that relay which showed low SER in Nakagami fading channel at low SNR and at high SNR, low SER is observed in Rician fading channel. While along with cooperation, Rayleigh is again following the footprints of direct path communication of Rician fading channel.



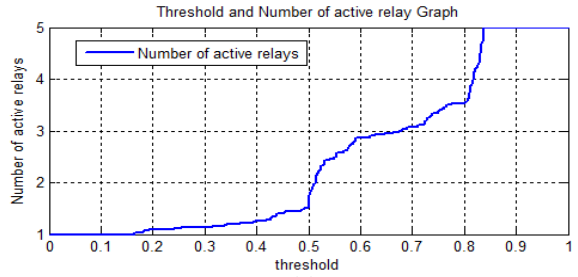
(a)



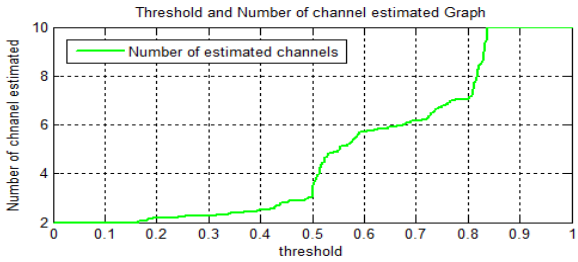
(b)

Fig. 5. SER analysis using Rayleigh, Rician and Nakagami fading channels using *fr* = 0.7 (a) *Algorithm-I* (b) *Algorithm-II*

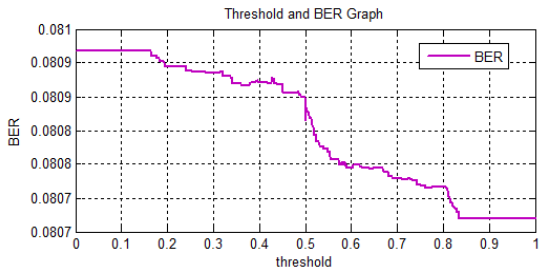
Fig. 6. shows the relation of threshold (f_r) with number of channel estimations, number of active relays and SER. This relationship is evaluated in Rayleigh fading channel using the SNR value equal to 1 and is only applicable for algorithm-2. Our simulation result proved that number of channel estimations and number of active relays (power consumption) increases as the threshold (f_r) increases. Fig. 6.(a) shows increase in the average number of active relays as threshold (f_r) increases reason for which is the fact that with increase in threshold, the algorithm will evaluate more and more relays to find the fittest one. Fig. 6.(b) shows the average number of channels estimated for relay selection and threshold (f_r) curve with the same trends and reasons as was for Fig. 6.(a) discussed above. Fig. 6.(c) shows a decrease in the average SER as the threshold (f_r) increases as the fittest on the relays will be selected with the increase in threshold.



(a)



(b)



(c)

Fig. 6. In Raleigh fadding enviroment curve of threshold vs (a) Active-relays (b) number of channel estimations (c) BER

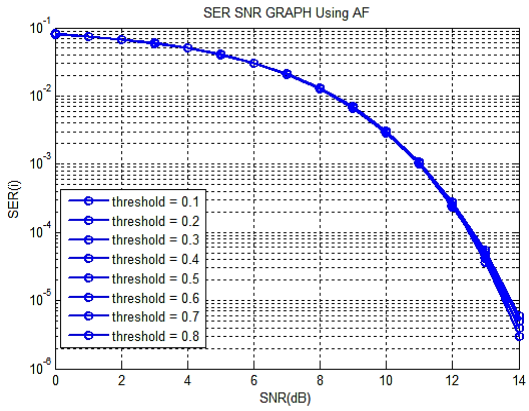


Fig. 7. BER analysis of *algorithm-II* in Rayleigh fading channel using different value of threshold

As clear from the figures that if we want low power consumption, less channel estimation (low load, less processing time for relay selection), system performance will become low in term of SER. To analyze the degradation of system performance in term of BER, performance using Rayleigh fading channel for different values of threshold (f_r) is evaluated. It is proved through our simulation result that the SER curves for different values of threshold (f_r) have very minute difference and almost equal at low SNR values. Fig. 7. shows the BER curve for different values of threshold (f_r).

VI. Conclusion and Future Work

In this paper, fuzzy logic based relay selection schemes are presented using three input fuzzy parameters i.e. SNR, cooperative gain and channel gain. In second proposed algorithm, threshold is used in order to minimize the power consumption and channel estimation required for relay selection. Simulation result showed that number of active relays and load of channel estimation is decreased by using the threshold. And also number of active relays are reduced so conserving power. SER is poorly effected at high SNR and almost not effected at low SNR. In future, this work can be extended by including the secrecy constraints in proposed relay selection schemes. Whereas, multi-hop communication in this system with multi-user detection and supplementary cooperation could increase the performance of the system manifolds while also increase in code rate is expected.

References

[1] Politis, C., et al., Cooperative networks for the future wireless world. IEEE Communications Magazine, 2004. 42(9): p. 70-79.

[2] Laneman, J.N., D.N. Tse, and G.W. Wornell, Cooperative diversity in wireless networks: Efficient protocols and outage behavior. IEEE Transactions on Information theory, 2004. 50(12): p. 3062-3080.

[3] Kharat, P. and J. Gavade, Cooperative communication: New trend in wireless communication. International Journal of Future Generation Communication and Networking, 2013. 6(5): p. 157-166.

[4] Laneman, J.N. and G.W. Wornell, Distributed space-time-coded protocols for exploiting cooperative diversity in wireless networks. IEEE Transactions on Information theory, 2003. 49(10): p. 2415-2425.

[5] Hasna, M.O. and M.-S. Alouini, End-to-end performance of transmission systems with relays over Rayleigh-fading channels. IEEE Transactions on Wireless Communications, 2003. 2(6): p. 1126-1131.

[6] Zhao, Y., R. Adve, and T.J. Lim. Improving amplify-and-forward relay networks: optimal power allocation versus selection. in 2006 IEEE International Symposium on Information Theory. 2006. IEEE.

[7] Zhao, Y., R. Adve, and T.J. Lim, Symbol error rate of selection amplify-and-forward relay systems. IEEE Communications Letters, 2006. 10(11): p. 757-759.

[8] Krikidis, I., et al., Amplify-and-forward with partial relay selection. IEEE Communications Letters, 2008. 12(4): p. 235-237.

[9] Duong, T.Q., V.N.Q. Bao, and H.-j. Zepernick, On the performance of selection decode-and-forward relay networks over Nakagami-m fading channels. IEEE Communications Letters, 2009. 13(3): p. 172-174.

[10] Su, W., A.K. Sadek, and K.R. Liu. SER performance analysis and optimum power allocation for decode-and-forward cooperation protocol in wireless networks. in IEEE Wireless Communications and Networking Conference, 2005. 2005. IEEE.

[11] Luo, J., et al., Decode-and-forward cooperative diversity with power allocation in wireless networks. IEEE transactions on wireless communications, 2007. 6(3): p. 793-799.

[12] Wang, C.-L. and S.-J. Syue. A geographic-based approach to relay selection for wireless ad hoc relay networks. in Vehicular Technology Conference, 2009. VTC Spring 2009. IEEE 69th. 2009. IEEE.

[13] Tanoli, U., et al., Performance analysis of cooperative networks with inter-relay communication over Nakagami-m and rician fading channels. International Journal on Multidisciplinary sciences, 2012. 3(4): p. 24-29.

[14] Tanoli, U., et al., Comparative analysis of fixed-gain relaying schemes for inter-relay communication over Nakagami-m fading channel. Sindh University Research Journal-SURJ (Science Series), 2013. 45(1).

[15] Madan, R., et al., Energy-efficient cooperative relaying over fading channels with simple relay selection. IEEE Transactions on Wireless Communications, 2008. 7(8): p. 3013-3025.

[16] Chen, Y., et al. Power-aware cooperative relay selection strategies in wireless ad hoc networks. in 2006 IEEE 17th International Symposium on Personal, Indoor and Mobile Radio Communications. 2006. IEEE.

[17] Wei, Y., et al. Energy-Saving Power Allocation Scheme for Relay Networks Based on Graphical Method of Classification. in International Conference on Human Centered Computing. 2016. Springer.

[18] Sun, L., et al., Cooperative communications with relay selection in wireless sensor networks. IEEE Transactions on Consumer Electronics, 2009. 55(2): p. 513-517.

[19] Li, Y., et al., Fair relay selection in decode-and-forward cooperation based on outage priority. Science China Information Sciences, 2013. 56(6): p. 1-10.

[20] Han, L. and J. Mu, Outage Probability of Opportunistic Decode-and-Forward Relaying over Correlated Shadowed Fading Channels. Wireless Personal Communications, 2016: p. 1-10.

[21] Zhu, Y. and H. Zheng, Understanding the impact of interference on collaborative relays. IEEE Transactions on Mobile Computing, 2008. 7(6): p. 724-736.

[22] Ju, M., K.-S. Hwang, and H.-K. Song, Relay selection of cooperative diversity networks with interference-limited destination. IEEE Transactions on Vehicular Technology, 2013. 62(9): p. 4658-4665.

[23] Shi, C., et al., Distributed interference-aware relay selection for IEEE 802.11 based cooperative networks. IET networks, 2012. 1(2): p. 84-90.

[24] Bletsas, A., et al., A simple cooperative diversity method based on network path selection. IEEE journal on Selected Areas in Communications, 2006. 24(3): p. 659-672.

[25] Fei, L., et al., Relay selection with outdated channel state information in cooperative communication systems. IET Communications, 2013. 7(14): p. 1557-1565.

[26] Bletsas, A., A. Lippnian, and D.P. Reed. A simple distributed method for relay selection in cooperative diversity wireless networks, based on reciprocity and channel measurements. in 2005 IEEE 61st Vehicular Technology Conference. 2005. IEEE.

[27] Wu, Q., X. Zhou, and S. Wang, Relay Selection Considering Successive Packets Transmission in Cooperative Communication Networks. CIT. Journal of Computing and Information Technology, 2014. 22(4): p. 217-226.

[28] Brante, G., et al., Distributed fuzzy logic-based relay selection algorithm for cooperative wireless sensor networks. IEEE sensors journal, 2013. 13(11): p. 4375-4386.

[29] Chen, M., T.C.-K. Liu, and X. Dong, Opportunistic multiple relay selection with outdated channel state information. IEEE Transactions on Vehicular Technology, 2012. 61(3): p. 1333-1345.

[30] Mamdani, E.H. and S. Assilian, An experiment in linguistic synthesis with a fuzzy logic controller. International journal of man-machine studies, 1975. 7(1): p. 1-13.

Wavelet-based Image Modelling for Compression Using Hidden Markov Model

Muhammad Usman Riaz
MCS, National University of Science and Technology Islamabad, Pakistan

Imran Touqir
MCS, National University of Science and Technology Islamabad, Pakistan

Maham Haider
MCS, National University of Science and Technology Islamabad, Pakistan

***Abstract*—Statistical signal modeling using hidden Markov model is one of the techniques used for image compression. Wavelet based statistical signal models are impractical for most of the real time processing because they usually represent the wavelet coefficients as jointly Gaussian or independent to each other. In this paper, we build up an algorithm that succinctly characterizes the interdependencies of wavelet coefficients and their Non-Gaussian behavior especially for image compression. This is done by extracting the combine feature of hidden Markov model and Wavelet transformation that gives us comparatively better results. To estimate the parameter of wavelet based Hidden Markov model, an efficient expectation maximization algorithm is developed.**

Keywords—Hidden Markov model; Wavelet transformation; Compression; Expectation Maximization

I. INTRODUCTION

Wavelet transformation is the tool for statistical signal processing and image modeling often used in real time signals processing[1]. Due to strong coordination between wavelet coefficients, these models have complicated processing but performance is much better. Wavelet transformation has primary and secondary properties that are used for statistical signal processing and image modeling[2]. These wavelet based hidden Markov models have many real-time applications in Engineering and Medical field.

In statistical signal processing techniques, wavelet transformation consider wavelet coefficient as single scalar coefficient that gives powerful tools for image modeling, previously discussed in [3,4,5]. These techniques consider coefficients as independent to each other but such methods those exploit dependencies between coefficients give better results.

In this paper we have developed a wavelet based hidden Markov model that succinctly model the statistical dependencies of the wavelet coefficients that are non-Gaussian in nature. Proposed algorithm exploits the statistical dependencies of the wavelet coefficients for better compression results. For this purpose we have used the combine properties of the wavelet transformation and Hidden Markov model.

A. Wavelet Transformation

Wavelet transformation is used to convert signals or images into its coefficients that contains complete information about signal or image [6]. Wavelet transformation converts the image into four subparts, first part contains approximation of the original image and remaining three parts contain the diagonal, horizontal and vertical coefficients information respectively [7].

Complete modeling of the image can be done with the help of wavelet coefficient and scaling coefficients. Primary properties (locality, multi-resolution and compression) of the wavelet transformation plays an important role in approximation of many real time signals [8]. With the help of locality each wavelet atom can be localized in time and frequency domain simultaneously. To analyze the wavelet atom at any scale we can use multi-resolution property of wavelet transformation [9].

Both locality and multi-resolution properties of wavelets facilitate us to match the large range of real time signals [10]. Most of the complicated signals can be approximated with small number of wavelet basis and scaling coefficients which will be discussed thoroughly in coming sections. We can conclude that all those statistical signal processing and image modeling methods that uses wavelet transformation are more beneficent as compared to those methods that uses only frequency-domain or time-domain information of the image/signal.

B. Statistical modelling for image

To approximate the wide range of coefficients we can use probability model based on wavelet transformation that is flexible, rich and tractable[11].

Previously wavelet coefficients modeled as non-Gaussian or jointly Gaussian[12,13] but statistically independent to each other [14]. To capture the linear dependencies of the wavelet coefficients, jointly Gaussian model is used. Histogram of wavelet coefficients density is peaky at zero indexes and heavily tailed than the histogram of the typical Gaussian distributions[15]. Complete decorrelation of wavelet coefficients is impossible, residual dependencies are always present between the wavelet coefficients after wavelet transformation[16]. Non Gaussian Models that doesn't exploit the complete statistical dependencies of wavelet coefficients during modeling are unrealistic to work with.

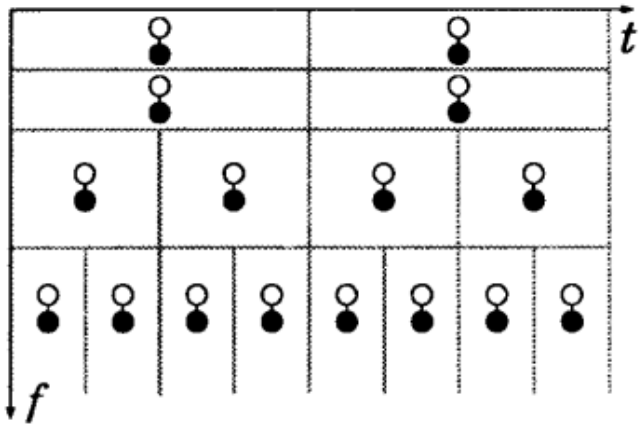


Fig. 1. Independent Mixture Model, each black and white node represent the continuous wavelet coefficients C_i and mixture state variable Q_i for C_i respectively

To overcome these problems, secondary properties of the wavelet transformation is helpful[17,18]. Clustering(one of the wavelet transformation property) suggests that for any high/low energy wavelet coefficient, neighboring coefficient will also be high/low energy coefficients[19]. According to persistence, values of the wavelet coefficients propagate across the scale. We will use these properties in the statistical modeling of the image.

To completely characterize the statistical dependencies of all wavelet coefficients we need to model the jointly probability density function that would consider all the dependencies of the wavelet coefficients but it is intractable and impossible to apply on real time images[20]. Conversely, modeling the wavelet coefficients without exploiting dependencies of the wavelet coefficients is simple and easy to implement but would not consider the inter-coefficients dependencies[21]. For better performance we need to make a balance between these two extremes.

We developed a wavelet based hidden Markov model that completely characterizes the probability structure of the wavelet coefficients. We model the marginal probability of each coefficient as mixture density with state variable to match the non-Gaussian nature of the wavelet coefficients as shown in figure 1. We introduce the Markovian dependencies between the hidden state variables to characterize the key dependencies between the wavelet coefficients. These dependencies are illustrated in figure 1.

Hidden Markov model is one of the Probabilistic graphical model used in statistical signal and image modelling. In this paper we used three probabilistic graph models with state-to-state connectivity as shown in figure 2. Independent mixture model ignores statistical dependence of wavelet coefficient and leave the variable states unconnected. Within each scale, to connect the state variable horizontally, we used the hidden Markov chain model. To connect the state variable vertically across the scale hidden Markov tree model is used. Together these three models represent the wavelet based hidden Markov Model.

Remaining distribution of the paper is as: Section 2 discusses the definitions and notation that will be used in mathematical modeling. Section 3 explains the statistical image modeling using hidden Markov model. Section 4 will elaborate the EM Algorithm and training of the proposed algorithm. Application of proposed framework will be discussed in section 5 along with experimental results. At the end section 6 will conclude the complete paper and suggest the future work.

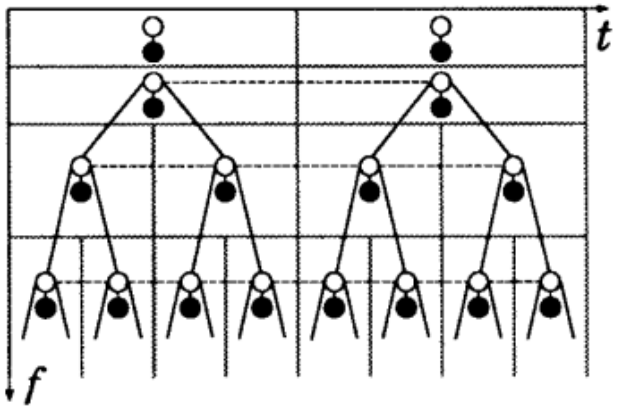


Fig. 2. Horizontal dashed line represent the connection capitulate HMC model and vertical solid line represent the connection capitulate HMT model respectively

II. NOTATIONS

Wavelet based hidden Markov model can completely characterize with the help of trees and graphs. $\{n_1, n_2, n_3, n_4 \ldots \ldots n_N\}$ used to represent the graph nodes. Term "connection" used to link the nodes. Ancestors are those nodes that are present on the path going from v_i to the root node. Descendants are those nodes that are present in the path from v_i going away from root node. $n_P(i)$ is parent of n_1 node if it is immediate ancestor. To denote the children of n_i we will represent it as $\{n_j\}\, j \in c(i)$. A node has many children but have only one parent. In case of binary tree one parent has only two immediate children. We will represent the c_i^k, i^{th} wavelet coefficient from the k^{th} tree. In this paper capital letters will be used for random variables (R.V) and small letters will be used as an observed value of that R.V. Probability mass function of discrete random variable "Q" denoted as $p_Q(q)$ and $f_C(c)$ as probability density function of continuous R.V "C".

III. WAVELET-DOMAIN PROBABILITY MODELS

Main problem is to exploit the key dependence of the wavelet coefficients during modeling of wavelet based hidden Markov model by considering the wavelet coefficients that follows non- Gaussian distribution. This modeling is done in two steps. Initiate with simple statistical model with the assumption that all wavelet coefficients are independent to each other based on the fact that wavelet transform de-correlate many of the real time signal's coefficients. Then enlarge this model using Markovian structure to exploit the residual dependencies of the wavelet coefficients. Here hidden Markov model is helpful that uses the state of the wavelet coefficients instead of values of the wavelet coefficients. Both first order Markovian dependencies and marginal Gaussian mixture provides to the implementation this model practically.

A. Wavelet Transformation

Primary properties of the wavelet transformation allow us to model each wavelet coefficient as high or low state. High state represents all those components carrying significant amount of information and low state is for less energy wavelet components. So, for each wavelet coefficient we have two states. 0-mean density and high variance is for high state whereas, 0-mean density and low variance for low state wavelet coefficients. Each wavelet coefficient is modeled as a 2-state mixture model illustrated in fig 3.

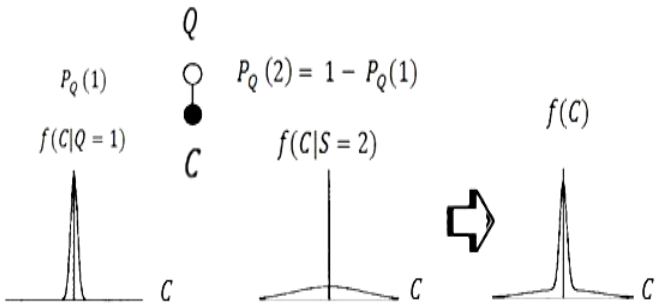


Fig. 3. Two-state, 0-mean GMM. White node represent the state variable and black node is for random variable.$\mathbf{p_Q(1)}$ Represent low variance (state 0) and $\mathbf{p_Q(2)}$ represent high variance (state 1)

Now we can completely characterize the 2-state, 0-mean with probability mass function of $p_Q(1)$ and $p_Q(2)$. Mostly Q (state variable) is hidden while C (value of coefficient) is observed.

For simplicity we will focus only on 2-state model. Consistent with $p_Q(q)$(probability mass function) "Q" can have only two values.

The probability density function of W is:

$$f_C(c) = \sum_{m=1}^{M} p_Q(m) f_{c|Q}(c|Q=m) \quad (1)$$

However, as "Q" is a random state variable and its value is hidden to us so probability density function will be non-Gaussian and overall wavelet coefficient has non-Gaussian density function even though wavelet coefficients are conditionally Gaussian. Normally scaling coefficients has non-zero variance so it is inappropriate to use 2-state, 0-mean Gaussian mixture model. Midway is to use Gaussian mixture model but with mixing densities that have non-zero means.

Wavelet transformation de-correlates the wavelet coefficients of many real time signals. So for statistical modeling of wavelet coefficients, independent Gaussian mixture model gives considerable improvement over deterministic signal. We have modeled one wavelet coefficient using 2-state Gaussian mixture model so it looks logical to use it for all the wavelet coefficients. Preferably, we need such a probabilistic model that collectively consider the probability density function of wavelet coefficients and exploit the statistical dependencies of the wavelet coefficients. After developing the Gaussian model for one coefficient, we extend the Gaussian mixture model to two coefficients with the help of jointly Gaussian mixture model.

Persistence and clustering recommended that if any wavelet coefficient is in state one (high state) then most probably its neighboring wavelet coefficient will also be in high state. Two such wavelet coefficients that are in neighbor can be modeled using Gaussian Mixture model variables that are in independent state. This simple modeling ensures the modeling of overall wavelet coefficients using the same method.

B. Wavelet Transformation Based Graph Models

To make a connection between states and wavelet coefficients, primary properties of wavelet transform are helpful. To represent the dependencies of wavelet coefficients there is link connected between variables. We have three ways to connect the dots and will be discussed latter in this section. To make horizontal and vertical connection between Q_i "state variable", HMT model and HMC model are used respectively. Transition probability gives information about the probability of transition from one state to the other. We can model dependencies of state variable using Hidden Markov model.

Parameter of hidden Markov model is defined as:

- $p_{Q_1}(m)$, Probability mass function for Q_1 (root node).
- $\epsilon^{mr}_{i,p(i)} = p_{Q_i|QS_{p(i)}}[m|Q_{p(i)} = e]$, conditional probability of $Q_i$given $Q_p(i)$ is in the state of e.
- $\mu_{i,m}$& $\sigma^2_{i,m}$ represent the variance and mean.

Above parameters are collectively called the parameter of the model and represented by "θ". Remember that for this paper we assume that we have two states with 0-mean.

$$f_{C_i}(c_i|\{C_j\}_{i\neq j}, \{Q_i = q_j\}_{i\neq j}, Q_i = q_j = f_{C_i}(c_i|Q_i = q_i) \quad (2)$$

As state of wavelet coefficients are hidden to us so wavelet based Hidden Markov model does not rely only on the wavelet coefficients of Markov structure. Suppose $J(i)$ represents the scale of wavelet coefficient (C_i and Q_i).

$$f_{C_i}(c_i|\{C_l\}_{J(t)>J(i)}) \neq f_{c_i}(c_i|C_{(i)}) \quad (3)$$

Generally, wavelet coefficients are not Markov. Wavelet state variables owe Markov nature that is why wavelet based HMM is efficient for modeling the wavelet coefficients.

IV. Expectation Maximization Algorithm

To estimate the model parameter "θ" of the wavelet based HMM, we need to train the data that consists of wavelet coefficients "C" of the image. These parameters include probabilities of mixture state, mean "$\mu_{i,m}$" and variance "$\sigma^2_{i,m}$" of the GMM. We then apply the maximum likelihood principle to find the parameters of the proposed model. Direct estimation using likelihood principle is hard to estimate. Because meaning of the estimation is to find the hidden state "Q" of the wavelet coefficients "C", means and variance of the Gaussian Mixture model. Expectation maximization is an iterative algorithm that is used to collectively find the model parameters "θ" and probabilities for the unobserved states "Q". Expectation maximization is also known as Baum Welch algorithm in the domain of hidden Markov model. Details of expectation maximization steps in the case of Hidden Markov chain and independent mixture model will be discuss in coming section.

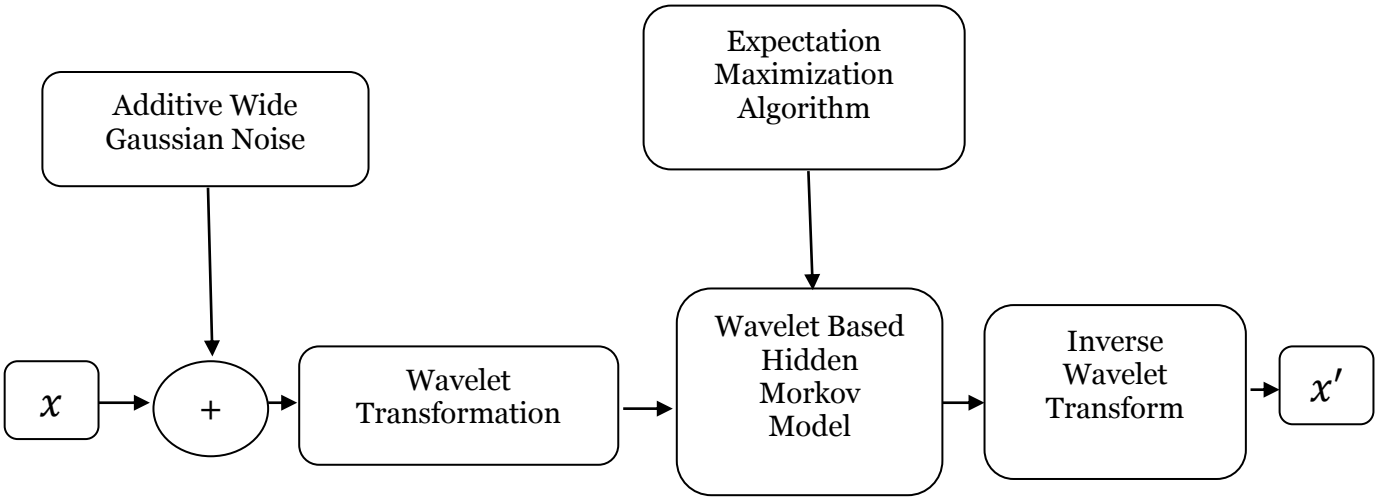


Fig. 4. Block diagram of the proposed wavelet based Hidden Markov Model

Figure 4 illustrate the complete functionality of algorithm graphically that is used to estimate the model parameters of the wavelet based hidden Markov model.

A. Expectation Maximization algorithm for training

In EM algorithm, "Training" is used for fitting the model parameters to the Wavelet based hidden Markov model. Basically training is used to avoid the "over-fitting". Our objective is to maximize the $\ln f(c|q)$ log function of wavelet coefficients "c" given hidden state variable "q" for this purpose expectation maximization algorithm performs two steps. In expectation step (called E-Step) algorithm, find the $E_S[\ln f(c, Q|\theta)|c,\theta^i]$ and in next step algorithm, maximize what it found in E-Step. Convergence will direct the algorithm in right direction if initialization is correct. Steps of the algorithm are given below:

Selection of a model parameter θ^0

Counter set to be zero, $i = 0$ (initialization)

- E step (expectation step)

Find $p(S|w, \theta^i)$, used in $E_S[\ln f(c, Q|\theta)|\ c\ ,\ \theta^i]$ maximization.

- M step (maximization step)

Set $\theta^{i+1} = \arg\max_\theta E_Q[\ln f(c, Q|\theta)|\ ,\ \theta^i]$

- update$i = i + 1$.

If it converges, then stop the iteration

Otherwise start again from step 1.

Initialization is important in EM algorithm. If we intelligently initialize the algorithm the complexity will reduce and it will start converging in few steps but in case if we randomly initialize, then it might be possible that after few iterations the complexity of algorithm exponentially increase and it will start diverging. To reduce the complexity of the algorithm we need to initialize the algorithm intelligently.

V. Experimental Results

Although this model works for statistical image processing in numerous real time signals. In this paper we have developed a wavelet based statistical model for image compression. We have compared the compression performance of our model by using different types of wavelets. Proposed model has a substantial improvement over previous models. Our results demonstrate the performance of the wavelet based Hidden Markov model in image compression.

TABLE I. Percentage of Compression Ratio and PSNR Image Name: Jelly Fish, Image Size 128x128, Level 1

S.No.	Wavelet (Level 1)	PSNR (db)	Compression Ratio
1	Bior 3.1	52.85	70.59%
2	Bior 3.3	53.70	70.14%
3	Bior 3.5	53.92	70.06%
4	Bior 3.7	53.98	70.08%
5	Bior 3.9	54.00	70.08%
6	Bior 4.4	56.82	62.28%
7	Bior 5.5	58.07	69.56%
8	Bior 6.8	56.69	72.66%

TABLE II. PERCENTAGE OF COMPRESSION RATIO AND MEAN SQUARE ERRORS IMAGE NAME WOOD STATUE, SIZE 256*256, LEVEL

S.No.	Wavelet type	MSE (db)	Compression Ratio
1	Haar	0.1201	65.85%
2	Sym2	0.1368	63.36%
3	Sym 3	0.1364	61.83%
4	Sym 4	0.1382	61.04%
5	Sym 6	0.1355	60.54%
6	Sym 8	0.1343	60.36%
7	Db1	0.1201	65.85%
8	Db 2	0.1368	63.36%
9	Db 3	0.1364	61.83%
10	Db 4	0.1306	59.31%
11	Db 6	0.1339	61.59%
12	Db 9	0.1340	62.20%
13	Coif1	0.1355	62.01%
14	Coif 2	0.1315	60.30%
15	Coif 3	0.1292	59.60%
16	Coif 4	0.1318	59.96%
17	Coif 5	0.1296	59.63%
18	Dmey1	0.1310	61.06%
19	Dmey 2	0.1110	68.52%
20	Rbio1	0.1123	73.96%
21	Rrbio 2	0.08184	77.30%
22	Bior1	0.1201	65.86%
23	Bior 2	0.1315	66.75%
24	Bior 3	0.1379	66.20%
25	Bior 4	0.1644	58.46%
26	Bior 5	0.1379	58.38%
27	Bior 6	0.1289	59.83%
28	Bior 7	0.1403	56.96%

A. *Objective analysis for image (2d) compression using wavelet domain hmm*

To achieve the image compression, we apply the different wavelets on proposed algorithm. We had applied 28 different wavelets and observe behavior of proposed algorithm. From Table 2, we can analyze that as the wavelet type changes, the compression ratio and mean square value also changes. Results illustrate that compression ratio depends of proposed algorithm depends on image characteristics as well as on the wavelet type used. When we apply dmey1 wavelet on the wood statue image, size 256x256 it gives 61.06% compression with 0.1310db mean square error (serial number 18 in table 2) and by changing the wavelet types we found different compression ratio as well as mean square error as shown in Table 2. "rbio 2" is the best wavelet for this test image because it provides 77.30% compression ratio with 0.08184db mean square error (serial number 21 in table 2).

Table 3 represent the comparison between different wavelet of same type "Brior" by comparting their means square error and compression ratio. We have found "Bior 6.8" best among all that gives 0.1392db mean square error and 72.66% compression ratio and "Bior 3.1" worst that give 0.337db mean square error and 70.59% compression ratio.

TABLE III. PERCENTAGE OF COMPRESSION RATIO AND MEAN SQUARE ERRORS IMAGE NAME: JELLY FISH, IMAGE SIZE 128X128, LEVEL 1

S.No.	Wavelet type	MSE (db)	Compression Ratio
1	Bior 3.1	0.3377	70.59%
2	Bior 3.3	0.2771	70.14%
3	Bior 3.5	0.2638	70.06%
4	Bior 3.7	0.2603	70.08%
5	Bior 3.9	0.2591	70.08%
6	Bior 4.4	0.1353	62.28%
7	Bior 5.5	0.1013	69.56%
8	Bior 6.8	0.1392	72.66%

B. *Subjective analysis for image (2d) compression using wavelet domain hmm*

The proposed algorithm is implemented and tested over the wide range of grayscale and colored images. The natural test images used are wood horse, Persons, Wood Statue, Mask, Facets, Laure, Catherine, Wood Statue, Arms and Jelly Fish. The results for these images are given in figure 5.

The goal of subjective image comparisons is to determine the effects of the following on compression performance: compression by using different types of wavelets on grayscale and colored images.

Jelly fish, one of the most common test images in compression research, consists primarily of low frequency content. Table 3 lists the MSE results and Compression ratios using different types of wavelets for jelly fish. The "rbio 2" wavelet depicts the best MSE performance among all.

VI. CONCLUSION AND FUTURE WORK

Wavelet based hidden Markov model is one of the model that is used for image modeling. These models are helpful in statistical modeling of wavelet coefficients of images that succinctly models the coefficients that don't follow the Gaussian distributions.

We have developed the wavelet based hidden markov model for statistical image modeling of the wavelet coefficients for compression. This model allows us to exploit the interdependencies of the wavelet coefficients and consider the entire coefficients during modeling that does not follow the Gaussian distribution. Mostly statistical methods model the wavelet coefficients as jointly Gaussian or independent to each other so, these models exploit less information about image characteristics. For parameter estimation of wavelet based hidden markov model, an efficient expectation maximization algorithm is developed.

Proposed approach gives the encouraging compression results in image compression domain, further research should be directed to multidimensional wavelet domain Hidden Markov Models as in this paper only statistical image modeling for 2-states is discussed.

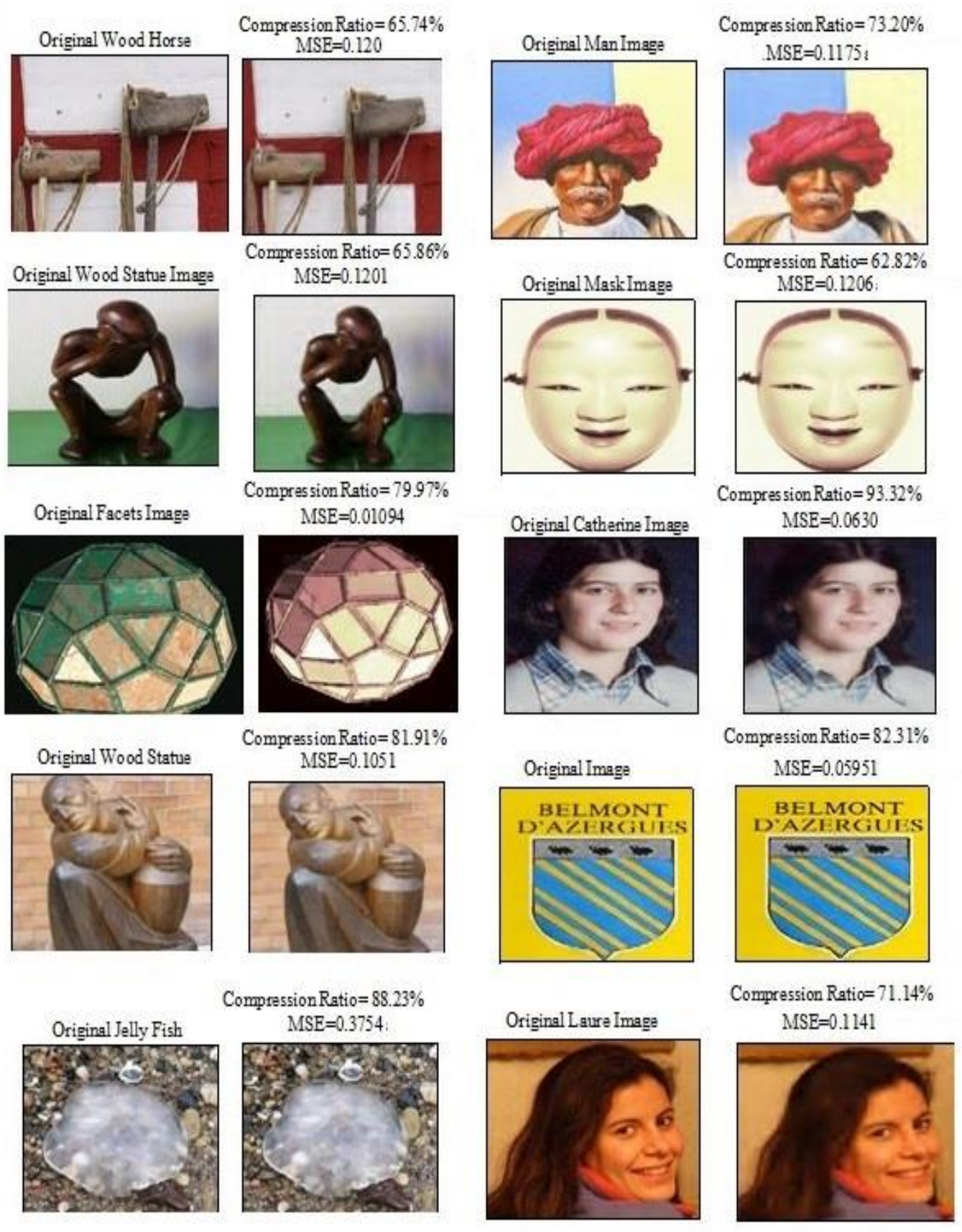


Fig. 5. Comparison between compression ratio and measn square error of differents images

REFERENCES

[1] Camps-Valls, G., Tuia, D., Bruzzone, L., and Atli Benediktsson, J. Advances in hyperspectral image classification: Earth monitoring with statistical learning methods. IEEE Signal Proc. Mag. 31, 1 (Jan. 2014), 45, 54.

[2] M. A. Losada, G. Tohumoglu, D. Fraile, and A. Artes, "Multi-iteration wavelet zerotree coding for image compression," Sci. Signal Process. vol. 80, pp. 1281–1287, 2000.

[3] M. S. Crouse and R. G. Baraniuk, "Contextual hidden Markov models for wavelet-domain signal processing," in Proc. 31st Asilomar Conf. Sig nals, Syst., Comput., Nov. 1997.

[4] H. Choi and R. Baraniuk, "Image segmentation using wavelet-domain classification," in Proc. SPIE, vol. 3816, Denver, CO, July 1999.

[5] Gilbert Strang and Truong Nguyen, Wavelets and Filter Banks, Wellesley-Cambridge Press, 1997.

[6] Y. U. Khan and J. Gotman, "Wavelet based automatic seizure detection in intracerebral electroencephalogram," Clin. Neurophysiol. vol.114, pp. 898–908, 2003

[7] K. R. Rao and P. Yip, Discrete Cosine Transform: Algorithms, Advantages, Applications. New York: Academic, 1990.

[8] J. M. Shapiro, "Embedded image coding using zerotrees of wavelet coefficients," IEEE Trans. Signal Process., vol. 41, no. 12, pp. 3445–3462, Dec. 1993.

[9] Kentaro Kinebuchi, signal interpolation using Wavelet based hidden Markov trees, 2000

[10] Yaonan Wang, Xiaoping Ma, Application of MultiScale Hidden Markov modeling wavelet coefficients to fMRI activation detection. (Report): An article from: Journal of Mathematics and Statistics.

[11] M. Alam, C.A. Rahman, W. Badawy, G. Jullien, Efficient Distributed Arithmetic Based DWT Architecture for Multimedia Applications, Proceedings of the 3rd IEEE International Workshop on System-on-Chip for Real-Time Applications, pages 333 -336, June 2003.

[12] M.Nibouche, A.Bouridane and O.Nibouche, A Framework for A Wavelet-Based High Level Environment, the 8th IEEE International Conference on Electronics, Circuits and Systems (ICECS), pages 429 - 432, vol.1, Sept. 2001.

[13] Boashash, M. Mesbah, and P. Coldtiz, "Time-Frequency Detection of EEG Abnormalities," in Time-Frequency Signal Analysis and Processing: A Comprehensive Reference, B. Boashash, Ed. Oxford, U.K.: Elsevier, 2003, pp. 663–670.

[14] G. K. Wallace, "The JPEG still-picture compression standard," Commun. ACM, vol. 34, pp. 30–44, Apr. 1991.

[15] Uwe Meyer-Baese, Digital Signal Processing with Field Programmable Gate Arrays, Springer-Verlag, 2001.

[16] Robert D. Turney, Chris Dick, and Ali M. Reza, Multirate Filters and Wavelets: From Theory to Implementation, Xilinx Inc.

[17] Spiliotopoulos, N.D. Zervas, C.E. Androulidakis, G. Anagnostopoulos, S. Theoharis, Quantizing the 9/7Daubechies Filter Coefficients for 2D DWT VLSI Implementations, 14th International Conference on Digital Signal Processing, pages 227 -231, vol.1, July 2002.

[18] J.Ramirez, A. Garcia, U. Meyer-Baese, F. Taylor, P.G. Fernendez, A. Lloris, Design of RNS-Based Distributed Arithmetic DWT Filterbanks, Proceedings of 2001 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), pages 1193 -1196, vol.2, May 2001.

[19] Mathematical Methods and Algorithms for Signal Processing, Todd K. Moon, Wynn C. Stirling, Prentice Hall, 2000

[20] Bochner S., Chandrasekharan K. (1949), Fourier Transforms, Princeton University Press Wavelet Transforms | A Quick Study, Ivan W. Selesnick, Polytechnic University,Brooklyn, NY,September 27, 2007

[21] Donoho, D.L.; I.M. Johnstone (1994), "Ideal spatial adaptation by wavelet shrinkage," Biometrika, Vol. 81, pp. 425–455.

Image De-Noising and Compression Using Statistical based Thresholding in 2-D Discrete Wavelet Transform

Qazi Mazhar
Military College of Signals,
National University of Sciences and Technology
Rawalpindi, Pakistan

Imran Touqir
Military College of Signals,
National University of Sciences and Technology
Rawalpindi, Pakistan

Adil Masood Siddique
Military College of Signals,
National University of Sciences and Technology
Rawalpindi, Pakistan

Adnan Ahmad Khan
Military College of Signals,
National University of Sciences and Technology
Islamabad, Pakistan

***Abstract*—Images are very good information carriers but they depart from their original condition during transmission and are corrupted by different kind of noise. The purpose is to remove the noisy coefficients such that minimum amount of information is lost and maximum amount of noise is suppressed or reduced. We considered Generalized Gaussian distribution for modeling of noise. In the proposed technique, statistical thresholding methods are used for the estimation of threshold value while Bi-orthogonal wavelet has been envisaged for image decomposition and reconstruction. A qualitative and quantitative analysis of thresholding methods on different images shows significant results for statistical thresholding methods based on objective and subjective quality as compared to other de-noising methods.**

Keywords*—Wavelet Thresholding; statistical Thresholding Image De-noising; Image Compression; Wavelet Sub-band Thresholding

I. INTRODUCTION

In the era of this digital world the use of digital images is greatly increased. Digital images are used in satellite, medical, radar, computer vision and pattern recognition. All these images are in digital form and noise is introduced in it during transmission, acquisition and processing. These images need to be de-noised before it is used in some kind of application. The goal of digital image de-noising is to restore the original image from a noise contaminated image and to preserve the important features of the image during the dropping the noisy coefficients. Now in [1] spatial domain filtering is used for the purpose of de-noising and in [2] transformed domain is used for image de-noising which shows improved results than spatial domain filtering. In transformed domain filtering, wavelets has superior results in image de-noising because it has some useful properties i.e. multi resolution analysis (MRA) and energy compaction. Instead of using spatial domain and Fourier domain the trend goes towards the wavelet transform domain. In wavelet domain the study shows that large coefficients of the images contains important features of the image and small coefficients mostly contains noise. Thresholding is an easy way to drop small coefficients and the noise will be removed efficiently. Thresholding used in the proposed method.

In the past few years a large amount of literature emerged on signal de-noising and comparison using different wavelet transform. In 1990's wavelet has been widely used in many fields of applications containing statistics estimation solving mathematical differential equations, Density estimation, image de-noising and compression. In 1995, Dohono and Johnstone invented a method of wavelet shrinkage which shows good results for 1-D signal de-noising and inverse problem solving [3]. These methods failed to meet improve the removal of noise from images. In de-noising and compression many of the coefficient values are dropped which are below the threshold value. The selection of the threshold has a great impact on the output image. Various methods are used to set a suitable optimal threshold value for the image thresholding [4, 5] but still the suitable optimal threshold value is big problem. Dohono and Johnstone invented universal thresholding for the optimal threshold value selection. The method finds a threshold value globally which is high value which drops a lot of useful information. The best threshold value is still a problem and challenge for researchers.

The wavelet transform will give us the translated and shifted version of the input image. The wavelet transform has time-frequency localization property. The shifted versions are frequency sub-bands which is used for the reconstruction of wavelet. During reconstruction it can restore the fine details of the input image and delete the unwanted coefficients of the noisy image. Different wavelet families are used for decomposition and reconstruction. The most recent and useful wavelet family is bi-orthogonal wavelet version 6.8. The comparison of orthogonal to bi-orthogonal wavelet family shows that bi-orthogonal has superiority in digital image processing. The bi-orthogonal wavelet has equal orthogonally and symmetry [6].

The gray scale images are actually composed of red, green and blue (RGB) color image but is presented in gray scale. In this paper we have used gray scale images for de-noising. By using statistical thresholding methods for de-noising and compression using two dimensional (2-D) discrete wavelet transform (DWT). In order to remove the noise and to retain the important features wavelet thresholding method and scale de-noising method is used for image de-noising and compression. [7, 8, 9, 10] proposed thresholding methods for noise removal which are more effective and easy to use and widely implemented. However all the above techniques have many drawbacks such as they are non-adaptive, having artifacts and blur. In this paper statistical thresholding method is used which are more adaptive and based on the statistics of the image. We have used bi-orthogonal 6.8 wavelet family which shows improved results, high de-noising, compression and edge preserving.

The paper is organized in five sections. Section I is introduction. Section II is introduction to 2-D DWT. Section III is de-noising techniques and statistical thresholding methods. Section IV shows the results and simulated data. Section V is conclusion and future work.

II. Discrete Wavelet Transform

Let the image be represented by $\{fij, i,j = 1\ 2\ \ldots.N\}$where N is power of 2.consider it is corrupted by additive white Gaussian noise and one observes

$$gij = fij + nij\ i,j\ = 1\ 2\ \ldots N \quad (1)$$

Where {nij} are independently and identically distributed (iid) as normal Gaussian distribution of $N(0, \sigma^2)$ and is independent of fij. The goal is to remove the noise or de-noise the noisy contaminated image $\{ gij \}$ and to obtain the approximated or estimated version of $\{ fij \}$.

Now let $G = gij$ and $F = fij$ and $N = nij$. The capital letters represents the matrix representation of the image which is under consideration. Now this is the representation of a noisy image. After passing this image from wavelet transform W. the equation becomes $Y=W_g$ the wavelet transform of noisy image. Here W is the 2-D dyadic orthogonal wavelet transform operator. W_f and $V=W_n$ are the wavelet transform of the input image and the noise respectively. The readers are referred to [11, 12].for details of 2-D orthogonal wavelet transform. In the figure below 2-D DWT decomposition of input image occurs. It is very easy to label the sub-bands of the transform. The sub-bands are HH_k, HL_k, LH_k, and k=1, 2…..j are called details. HH is diagonal details, HL are horizontal details and LH are vertical details. The low pass filters LL is called approximations, this contains the approximated coefficients values of the image. Here k is the scale and j is the level of decomposition. The sub-band at k scale has $N/2^k$ x $N/2^k$size. Now the transform is orthogonal so the {Vij } is *iid* $N(0, \sigma^2)$.

The coefficients of the details sub band i.e. HHk, HLk, LHk are pass by wavelet thresholding method for finding the approximated or estimated output $\{\hat{f}ij\}$. The lowpass band i.e LL1 is called approximation at decomposition level 1. This section of the image is further decomposed into horizontal, vertical, and diagonal details. The coefficients of this approximation are kept. After passing the details from thresholding operator the estimated output is passed from inverse wavelet transform $\hat{f}$ =W^{-1} x where W^{-1} is the inverse wavelet transform operator.

III. Image De-Noising Techniques

The most investigated domain in image de-noising using wavelet transform is nonlinear thresholding methods. Wavelet transform domain shows sparse property and wavelet maps noise from image domain to wavelet domain thus, the energy of the image is concentrated in high coefficients while noise energy is mostly in low coefficients values. This principle enables the separation of image important features from noise [13]. Now the procedure in which small coefficient values are dropped in large coefficients values left is known as hard thresholding but the drawback it produces visual artifacts. This is because of the unsuccessful attempt of removing large coefficients values. To overcome this problem soft thresholding was introduced. In this method the coefficient values shrinks towards the threshold value T. Most of the wavelet literature is about finding in optimal threshold value. Which can be adaptive or non-adaptive to the image. In wavelet thresholding there are two types which are mostly used for image de-noising and compression. One is soft thresholding and the other is hard thresholding:

Soft thresholding: It is also called shrinkage function. It shrinks the coefficient towards the threshold value T. it is a smoothing operator.

$$D(U,\lambda) = \text{sgn}(U) * \max(0, |U| - \lambda) \quad (2)$$

Hard thresholding: It is the function which either keeps the coefficient or kills the coefficient value. The result of this technique have very sharp edge.

$$D(U,\lambda) = \begin{cases} U \text{ for all } |U| > \lambda \\ = 0 \text{ otherwise} \end{cases} \quad (3)$$

A. Universal Thresholding(Visu shrink)

In the wavelet de-noising literature the universal thresholding is the most widely used one. It is globally approached and can be formulated as follows:

$$\lambda_T = \sigma\sqrt{2logN} \quad (4)$$

Where N is the size of the image and σ is noise variance. The λ_T must above the max level but not too large. Too much large coefficients may not be averted with increase in N Length and the threshold also increases due to Gaussian distribution.

Universal thresholding does not require prior information exactly like the Bayesian thresholding. For smooth data like Dohono it may be applied easily and conveniently.

When the size of the input signal is so large that it approaches to infinity the universal thresholding is the best candidate in that scenario. Also it is a good approach for statistical smoothness whose asymptotic behavior is better the mean square error.

This approach is too much fast and easy. Its implementation is straight forward, however when implemented on an image it produce a de-noised image which lost enough information

B. Statistical Thresholding Method 1

In this method we find the mean of each detail sub-band 'μ'. The σ_y is the variance of the degraded image which can be find by robust median estimator

$$\sigma_y^{\,2} = [\text{median}(|\text{each sub} - \text{band}|)/0.6745] \quad (5)$$

The noisy coefficients are very small and the signal coefficient are very large contains the useful information of the image. After the decomposition of the image to N level. The coefficients of the detail sub-bands are stored an array. Those values whose are greater than $2\sigma_y$, $3\sigma_y$ are dropped and the other values are kept. i.e.

$$y > \begin{cases} 2\sigma_y,\ 3\sigma_y; x = 2\sigma_y \\ \text{Else } y = y \end{cases} \quad (6)$$

Finding the noise variance σ_n and threshold value, finally add the value with mean 'μ'

$$t = \sigma_n^{\,2}/\sigma_s^{\,2} \quad (7)$$

C. Satistical Thresholding Method 2

The Statistical Thresholding method is effective for images including Gaussian noise. The observation model is expressed as follows:

$$Y = X + N \quad (8)$$

Here Y is the wavelet transform of the degraded image, X is the wavelet transform of the original image, and V denotes the wavelet transform of the noise components following the Gaussian distribution N (0, $\sigma_n^{\,2}$). Here, since X and V are mutually independent, the variances $\sigma_y^{\,2}$, $\sigma_x^{\,2}$ and $\sigma_n^{\,2}$ of y, x and n are given by

$$\sigma_y^{\,2} = \sigma_x^{\,2} + \sigma_n^{\,2} \quad (9)$$

It has been shown that the noise variance can be estimated from the first decomposition level diagonal sub-band HH1 by the robust and accurate median estimator [4] by (5).

The variance of the sub-band of degraded image can be estimated as:

$$\sigma_y^{\,2} = 1/M \sum_{m=1}^{M} (A\,m)^2 \quad (10)$$

Where Am are the wavelet coefficients of sub-band under consideration, M is the total number of wavelet coefficient in that sub-band. The statistical thresholding method 2 technique performs soft thresholding, with adaptive data driven, sub-band and level dependent near optimal threshold given by

$$T = \begin{cases} \frac{\sigma_n^{\,2}}{\sigma_x^{\,2}} & if\ \sigma_n^{\,2} > \sigma_y^{\,2} \\ \max(Am) & otherwise \end{cases} \quad (11)$$

IV. Simulation Results

To evaluate the performance of above techniques we applied it on different images. Five images Lena, Barbara, house mcslibrary and cameraman are used as test images. All the five images are of size (512 x 512) are applied to the above techniques at different standard deviation levels σ = 15, 20,25,30,35. We investigated different wavelet families. Bi orthogonal 6.8 (Bior6.8) [6] wavelet has superior results. We applied bi orthogonal 6.8 (Bior6.8) wavelet and decomposition level five in our simulations.to check the performance we compared the results with hard thresholding, soft thresholding, visu shrink, statistical method 1 and statistical method 2 using Peak signal to noise ratio (PSNR)[16]

$$\text{PSNR} = 10\ \log 10_{10}\left(\frac{\left(\max(f(m,n))\right)^2}{\text{MSE}}\right) \quad (12)$$

In the above equation f(m, n) shows the input image. We are dealing here with gray scale images. So we have

$$\left(\max(f(m,n))\right) = 255 \quad (13)$$

Where MSE is the mean square error between the degraded image and original image formulated as below.

$$\text{MSE} = \sum_{MN} \frac{f(m,n) - \tilde{f}(m,n)}{M \times N} \quad (14)$$

The other parameter used for psych-visual comparison is structural similarity index (SSIM) [16] which shows the structural similarity of the two images. SSIM index is calculated between two images X and Y is formulated as:

$$SSIM(X,Y) = \frac{(2\mu_x\mu_y + C1)\ (2\sigma_{xy} + C2)}{(\mu_x + \mu_y + C1)\ (\sigma_x + \sigma_y + C2)} \quad (15)$$

Here

μ_x *is the average of x*

μ_y *is the average of y*

σ_x *is the variance of x*

σ_y *is the variance of y*

σ_{xy} *is the covariance of xy*

C1 and C2 are constants

For subjective analysis mean opinion score (MOS) is used from decades. MOS is a test that shows the human view about the quality of images. It is ranked from unacceptable to excellent in the number from 1 (worst) to 5 (best). It is the averaged value of many users opinion [14, 15]. The table is shown below:

TABLE I. Mean Opinion Score (MOS)

MOS	QUALITY
1	Unacceptable
2	Poor
3	Fair
4	Good
5	Excellent

The graph in figure 4. is the PSNR versus noise variance for the image Lena shows that the statistical method 1 and statistical methods 2 performs better than visu shrink, soft and hard thresholding. As we increase the noise variance the PSNR value decreases. The plot shows the PSNR value of the noisy image along with existing and proposed techniques. Soft thresholding has improved the PSNR value to a little extent. Hard thresholding improved the result of soft thresholding. Visu shrink has improved the value of PSNR. Statistical

thresholding method 1 and 2 both shows superior results than all other techniques.

Now the MSE (Mean square error) Plot of the image lena shows graph of noisy image whose values are very high. The lower the MSE values the greater noise is removed. The plot shows that the soft, hard and visu shrink has lowered MSE to some extent. The statistical thresholding method 1 has lowered the layer of MSE to bottom level. Statistical method 2 has crossed statistical method 1 in removing error form the image. All the method is shown in the graph below

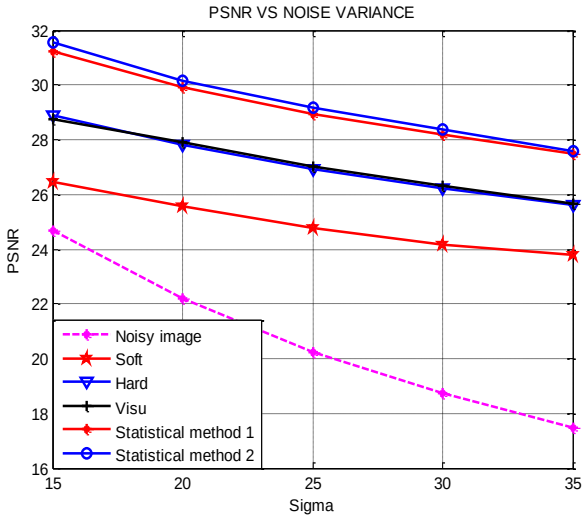


Fig. 1. Plot of PSNR vs Noise Variance for Sigma values=15, 20,25,30,35 for image Lena

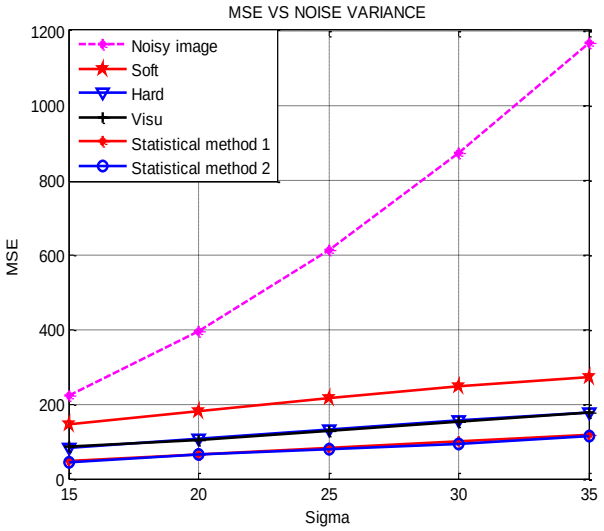


Fig. 2. Plot of MSE vs Noise Variance for Sigma values=15, 20,25,30,35 for image Lena

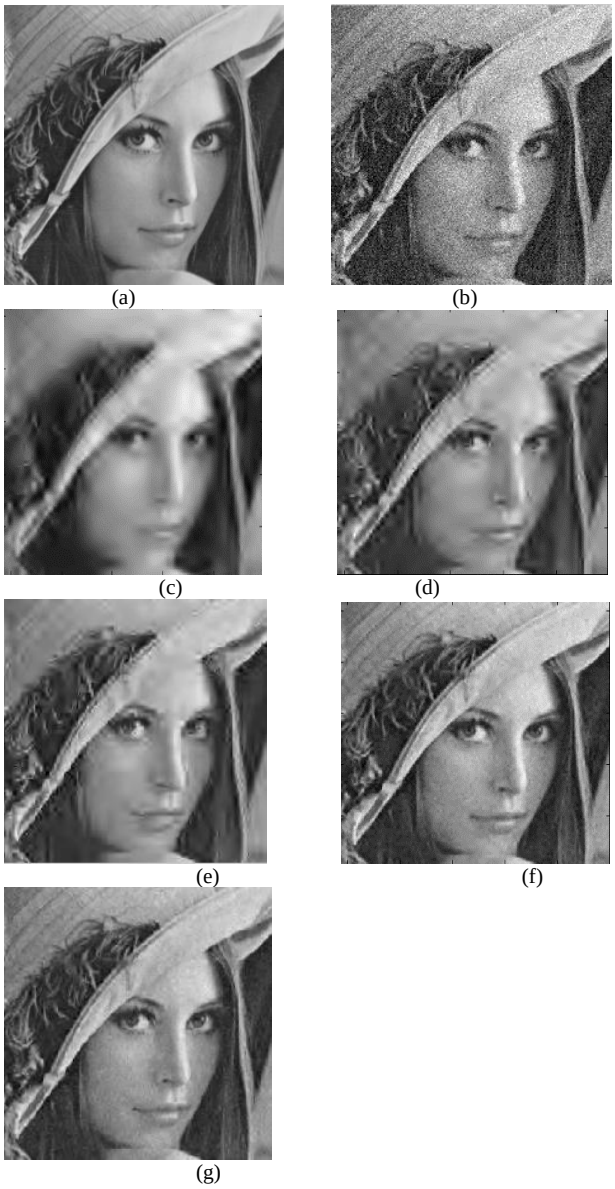


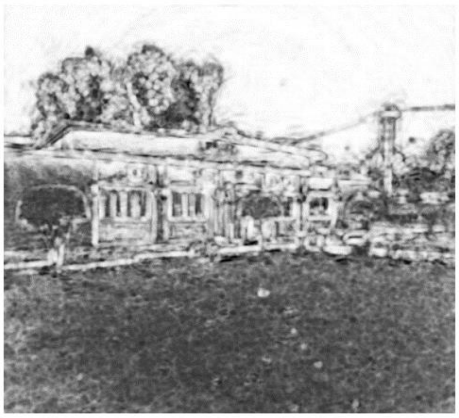
Fig. 3. (a) Original Lena image (b) Noisy Image with variance sigma =20 (c) Soft thresholding (d) Hard thresholding (e) visu shrink (f) Statistical thresholding method 1 (g) Statistical thresholding method 2

Visual quality of the images shows that soft thresholding has a blurred image but result is nearly a smoothed image as we know that soft thresholding gives us smoothed image. The result of hard thresholding is a sharp image having a large amount of artifacts. The visu shrink has improved result than hard thresholding. The statistical thresholding method 1 has more improved results than all previous method and removed artifacts and blur areas from the image. The statistical thresholding method 2 has improved the results of statistical method 2 and the visual quality is improved. The figure shows detail of all the methods.

The SSIM (Structural similarity index) shows us a mean value and similarity map of the original and de-noised image. As the mean value approaches to 1 the noise approaches to zero. The SSIM map and mean values of soft thresholding method is very small shows up to 50% similarity to original image. The hard thresholding method shows 55% similarity visu 56%, statistical method 1 61% and statistical method 2 shows 66% similarity of original ad de-noised image. The results of SSIM for the image of Military College of signals library MCS library is shown below.

(a)

(b)



(c)

(d)

(e)

(f)

Fig. 4. (a) Original image of MCS library SSIM map for (b) soft thresholding (c) hard thresholding (d) Visu shrink (e) Statistical method 1 (f) Statistical method 2

TABLE II. SSIM MEAN VALUE FOR DIFFERENT METHODS FOR IMAGE MCS LIBRARY FOR NOISE VARIANCE 20

Method	Mean value
Soft thresholding	0.5072
Hard thresholding	0.5503
Visu Shrink	0.5627
Statistical method 1	0.6142
Statistical method 2	0.6671

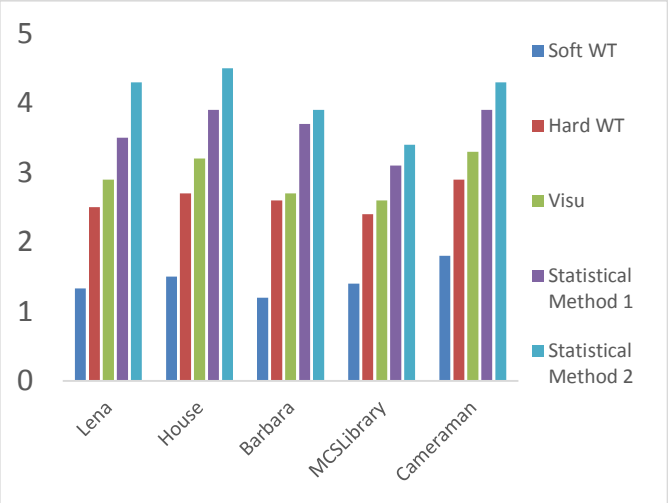


Fig. 5. Average value graph of mean opinion score for five images passed by five techniques

Fig. 5 shows an averaged value of MOS. The statistical thresholding method 2 has greatly minimized the noisy coefficients so the subjective quality for this technique is high while the other methods are arranged in the decreasing way showing his own image quality.

V. CONCLUSION

Image de-nosing and compression is the basic application of 2-D DWT. We have applied statistical thresholding methods on different images. We concluded that many thresholding methods used are not suited well for de-noising gray scale images. The results shows that for existing techniques many of the useful information of the image is lost which shows that degraded quality. The visual quality measured by MOS shows that existing image de-noising techniques produces blurriness and artifacts in the image. The statistical thresholding method 1 and statistical thresholding method 2 has largely improved PSNR values and MSE. The statistical thresholding methods also shows a higher structural similarity SSIM mean value. Statistical method 1 and statistical method 2 are adaptive and deals with each and every sub-band and coefficient value. The previous methods were level dependent thresholding methods but were non adaptive. Statistical thresholding methods are more effective due to adaptive and coefficient relevance judgment. In future work we will apply these thresholding methods on true color image.

VI. FUTURE WORK

In the future, the work done here can be extended in many directions. This work can be extended to analyze those images corrupted by salt and speckle noise, pepper noise and other noise models. The benefit of DWT is give visually a pleasing image and improves PSNR values. The work can be extended in such a manner to reduce the noise with no loss of actual information. Research can be done on different image to choose a suitable mother wavelet for de-noising and compression images effectively. The methods applied in this paper can also be extended to RGB images as well as video sequences.

ACKNOWLEDGMENT

The authors would like to thank editors and reviewers for their useful comment. This research was supported by Image processing center National University of engineering and technology (NUST) Islamabad, Pakistan.

REFERENCES

[1] A. Sharma and J. Singh, "Image de-noising using spatial domain filters: A quantitative study," Image and Signal Processing (CISP), 2013 6th International Congress on, Hangzhou, , pp. 293-298. 2013 doi: 10.1109/CISP.2013.6744005

[2] T. Shah, G. Shikkenawis and S. K. Mitra, "Epitome based transform domain Image De-noising," Advances in Pattern Recognition (ICAPR), 2015 Eighth International Conference on, Kolkata, , pp. 1-6. 2015 doi: 10.1109/ICAPR.2015.7050652

[3] Iain M.Johnstone David L Donoho. Adapting to smoothness via wavelet shrinkage.Journal of the Statistical Association,90(432):1200–1224, Dec 2007

[4] K. I. Kim, and Y. Kwon, Example-based Learning for Single-Image Super-Resolution and JPEG Artifact Removal. Technical Report No. TR-173, Max Planck Institute for Biological Cybernetics, August 2008.

[5] M. Mastriani y A. Giraldez, "Microarrays denoising via smoothing of coefficients in wavelet domain," WSEAS Transactions on Biology and Biomedicine, 2005

[6] M. Mastriani y A. Giraldez, "Fuzzy thresholding in wavelet domain for speckle reduction in Synthetic Aperture Radar images," ICGST International on Journal of Artificial Intelligence and Machine Learning, Volume 5, 2005

[7] M. Mastriani, "Denoising based on wavelets and deblurring via self-organizing map for Synthetic Aperture Radar images," ICGST International on Journal of Artificial Intelligence and Machine Learning, Volume 5, 2005

[8] M. Mastriani y A. Giraldez, "Kalman' Shrinkage for Wavelet-Based Despeckling of SAR Images," International Journal of Intelligent Technology, Volume 1, Number 3, pp.190-196, 2006.

[9] S. Grace Chang, Bin Yu and M. Vattereli. Spatially Adaptive Wavelet Thresholding with Context Modeling for Imaged noising. IEEE Transaction - Image Processing, volume 9, pp. 1522-1530. 2000.

[10] Maarten Janse. Noise Reduction by Wavelet Thresholding. Volume 161, Springer Verlag, United States of America, I edition. 2001.

[11] J. Chen, C. Tang, and J. Wang. Noise brush: interactive high quality image-noise separation. ACM Trans. Graphics, 28(5), 2009.

[12] Z. Wang, A. C. Bovik, H. R. Shiekh and E. P. Simoncelli,"Image quality assessment: From error visiability to structral similarity," IEEE Transaction on Image processing, vol 13, no. 4, pp.600-612, Apr. 2004

[13] Atidel Lahouhou, Emmanuel Viennet and Azeddine Beghdadi Selecting Low-level Features for Image Quality Assessment by Statistical Methods Journal of Computing and Information Technology - CIT 18, 2010,

Denoising in Wavelet Domain Using Probabilistic Graphical Models

Maham Haider
Military College of Signals
National University of Sciences and Technology, Islamabad, Pakistan

Imran Touqir
Military College of Signals
National University of Sciences and Technology, Islamabad, Pakistan

Muhammad Usman Riaz
Military College of Signals,
National University of Sciences and Technology, Islamabad, Pakistan

Adil Masood Siddiqui
Military College of Signals
National University of Sciences and Technology, Islamabad, Pakistan

***Abstract*—Denoising of real world images that are degraded by Gaussian noise is a long established problem in statistical signal processing. The existing models in time-frequency domain typically model the wavelet coefficients as either independent or jointly Gaussian. However, in the compression arena, techniques like denoising and detection, states the need for models to be non-Gaussian in nature. Probabilistic Graphical Models designed in time-frequency domain, serves the purpose for achieving denoising and compression with an improved performance. In this work, Hidden Markov Model (HMM) designed with 2D Discrete Wavelet Transform (DWT) is proposed. A comparative analysis of proposed method with different existing techniques: Wavelet based and curvelet based methods in Bayesian Network domain and Empirical Bayesian Approach using Hidden Markov Tree model for denoising has been presented. Results are compared in terms of PSNR and visual quality.**

Keywords*—Guassian Mixture Models (GMM); Hidden Markov Model (HMM); Discrete Wacelet Transform (DWT); Hidden Markov Tree (HMT)

I. Introduction

Removing noise from an image with less loss of information is referred as Image Denoising. Even the simplest acquisition, processing and transmission of image can subject it to unwanted noise. Denoising is required at every level whenever there is a question about dealing with images [1]. It finds its applications ranging from image enhancement in terms of highlighting some useful aspects of information from an image to achieving better quality of medical images (CT scan, Ultrasound etc).

There have been a number of techniques and algorithms designed in wavelet domain. All of them worked successfully because of primary properties that wavelet transform has that is; locality, multi-resolution and compression. In [2], a framework for signal denoising in time frequency analysis domain using Hidden Markov Model (HMM) has been proposed. A neighbouring coefficients thresholding method is applied in multi-wavelet framework for denoising in [3]. A non-parametric tree based model for joint statistics of wavelet coefficients has been discussed in [4]. To realize neighbouring dependency between wavelet coefficients across scales, a generalized Multivariate Gaussian distribution has been proposed in [5]. In [6], another tree model using hidden markov tree structure has been developed that refers to local parameterization for image denoising. In [7], wavelet shrinkage function is used to check neighbouring coefficients dependencies for image denoising. In [8], a denoising scheme has been realized that incorporates dependency of wavelet coefficients with three scale dependency. In [9], another image denoising technique based on neighbouring wavelet coefficients has been proposed.

Although there exists a number of denoising algorithms but there is still a need for improvement. Further research and study can result in an improved performance and better image quality. In this work, we have developed HMM based image denoising algorithm which is used in the context of 2D Gaussian Mixture Models (GMM) and 2D DWT. EM algorithm iteratively finds the maximum likelihood of a fundamental distribution from a given data set when the data is said to have some missing values. It is best suited in HMMs where the hidden states are not observed and the data is said to be missing. This modelling framework summarizes the nature of wavelet coefficients in a probabilistic way. This model finds its flexibility to the two main features; one is the Mixture Densities for dealing with the non-Gaussian nature of coefficients of wavelet transform by modelling them with hidden states as Mixture Distributions. The second feature is that of Probabilistic Graphs which models coefficients interdependencies and represent them in the form of a tree structure. Several local and standard images are put to test to check the performance parameters comparable to different other existing techniques for denoising. Results are shown both in terms of Peak Signal to Noise Ratio (PSNR) and the quality of denoised image.

II. Statistical Image Modelling Using Hidden Markov Tree Model

Wavelet transform is known to be favorable because of its properties such as non-Gaussianity, Clustering and Persistence [5], [7], [8]. Fig.1, shows 2D discrete wavelet transform

(DWT). In this figure the decomposition upto 3 levels is shown. HMM combined with these properties provides an attractive model for capturing the non-Gaussian parameters of the wavelet coefficients that are persistent across scales [2], [10]. The details of wavelet transform and wavelet domain HMM along with applications in statistical image processing can be found in detail in [1], [2], [11], [12], [13] and [14].

Fig. 1. Three level decomposition using 2D Discrete Wavelet Transform

For each wavelet coefficient magnitude $|c_{j,i}|$, we can associate a set of Hidden states ($S_{j,i}$) with it. Given $S_{j,i} = k$, the pdf will be Gaussian with mean and variance as $\mu_{j,k}$ and $\sigma^2_{j,k}$ respectively. The overall pmf will be;

$$f(c_{j,i}) = \sum_{k=1}^{K} P(S_{j,i} = k) f(c_{j,i} | S_{j,i} = k) \quad (2)$$

where $P(S_{j,i} = k)$ is the probability mass function (pmf) and $f(c_{j,i}|S_{j,i} = k)$ is the conditional pmf given by the following equation;

$$f(c_{j,i}|S_{j,i} = k) = \frac{1}{\sigma_{j,k}\sqrt{2\pi}} exp(-\frac{(c_{j,i} - \mu_{j,k})^2}{2\sigma^2_{j,k}}) \quad (3)$$

Owing to persistence across scale, state transition matrices A_t shows the parent → child state-to-state links between hidden states given as;

$$A_t = \begin{bmatrix} p_t^{m\to m} & p_t^{m\to n} \\ p_t^{n\to m} & p_t^{n\to n} \end{bmatrix} \quad (4)$$

where $p_t^{s\to s'}$ shows that given s' being the parent coefficient state, the child coefficient is in state s [15]. HMT model on the whole is specified by $P(S_o = k)$, pmf of the node S_o, a state transition matrix A_t and $\mu_{j,k}, \sigma^2_{j,k}$, means and variances, of the wavelet coefficient $c_{j,i}$ given $S_{j,i}$ being in the state k. All these parameters are combined together in a vector θ. The state transitions and variances are generally different for each wavelet coefficient. This can lead to a more complex HMT model. To reduce the computational complexity, a method referred to as tying within scale is implemented as discussed in [2].

H. Chipman has shown that GMM is capable of well approximating this non-Gaussian density in [16]. GMM has a generative model with a random variable Z and marginal distribution as $\sum_{k=1}^{m} \alpha_k N(c_{j,i} \mid \mu_{j,k}, C_{j,k})$ where $C_{j,k}$ represents the co-variance matrix of coefficients. A multidimensional GMM is referred to as HMT. To capture the non-Gaussianity of the wavelet coefficients which is referred to the Clustering property, it associates each of the wavelet coefficients magnitude $|c_{j,i}|$ with a hidden variable ($S_{j,i}$) that is said to be unobserved. We find few of the coefficients with large magnitudes and they tend to contain the maximum amount of information regarding the image. On the other hand, the coefficients with smaller magnitudes have the lesser amount of relevant information about the image but they tend to exist in large numbers. This leads to the simplest model with only two states; 'high' and 'low' referring to large and small magnitudes of wavelet coefficients respectively. Several results have proved this model to be simple in nature but effective in estimation point of view.

To capture the inter-scale dependencies between the wavelet coefficients, referring to the Persistence property [15], [17], GMM performs Markovian Chain across scales that are tree-structured as the magnitude of the coefficients is said to be only depending on the size of their respective parents alone. This means that the probability of a coefficient to be 'large' or 'small' is determined only by the magnitude of its parent. HMT in time-frequency domain was developed in [2] that connect the state variables vertically. For images it forms a structure in a quad-tree fashion that associates each hidden state variable as parent to its four children. This model satisfies both secondary properties that wavelet transform has such as Clustering and Persistence. HMT is applied to image processing in [11], [13] and [15].

III. Model Training Via EM Algorithm

In this section a denoising model has been discussed that uses EM algorithm to determine noise-free coefficients from noisy elements. Translating the problem in image denoising, there is a need to determine noise-free coefficient *c* from the noisy coefficients *y*. Estimate $c_{j,i}$ such that $y_{j,i} = c_{j,i} + n_{j,i}$, where $n_{j,i} = \mathcal{N}(0, \sigma_n^2)$. Generalizing it, this model refers to determining the vector denoted by θ. The following relation of expectation describes the sufficient statistics S_t of the model for variable X and the hidden state variable Z;

$$E_{\theta_o}(S_t(X,Z) \mid X = x) = E_\theta S_t(X,Z) \quad (5)$$

The conditional pmf of hidden states $S_{j,i}$ and its maximization is given by following expressions;

$$P(S_{j,i} = k \mid c_{j,i}, \theta') = \frac{P(S_o = k) g(c_{j,i}; 0, \sigma^2_{j,k})}{\sum_{l=0}^{1} P(S_o = l) g(c_{j,i}; 0, \sigma^2_{j,l})} \quad (6)$$

$$P(S_o = k) = \frac{1}{N_j} \sum_{i \in \mathbb{Z}^2} P(S_{j,i} = k \mid c_{j,i}, \theta') \quad (7)$$

The noise-free coefficients $c_{j,i}$ can be obtained from the following expression;

$$c_{j,i} = E[c_{j,i} \mid y, \theta] = \sum_{i \in \mathbb{Z}^2} P(S_o = k) \mid y, \theta) \frac{\sigma^2_{i,k}}{\sigma_n^2 + \sigma^2_{i,k}} y_i \quad (8)$$

Fig.2 shows the denoising scheme used in this research work. The proposed scheme for image denoising can be summarized as follows;

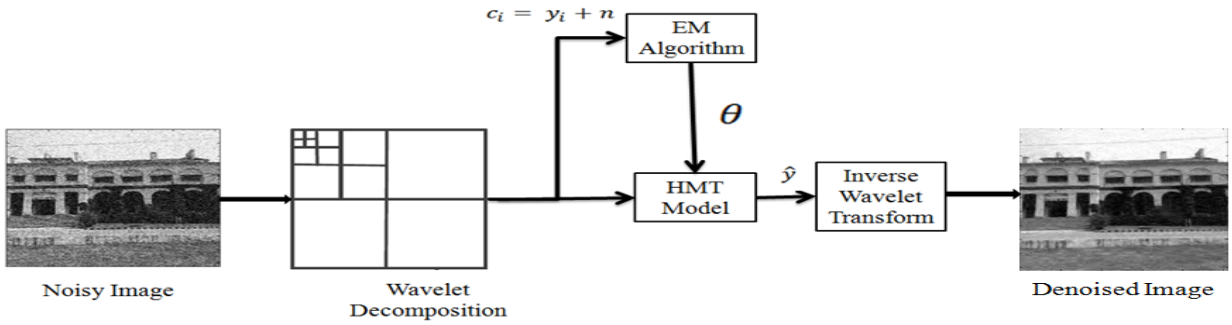


Fig. 2. Proposed Denoising Scheme

1) Add AWGN noise in the real-world image (NxN dimensions) after converting it into gray scale

2) Apply Daubechies-8 DWT on the image to divide it into three levels of DWT

3) Estimate the initial GMM parameters of each sub-band from the noisy coefficients.

4) Train the HMT model using EM algorithm with reference to tying within scales method

5) Apply inverse wavelet transform to estimate the noise-free coefficients to obtain the denoised image

IV. Experimental Results

We have tested standard as well as local gray scale images in order to compare this scheme with several other denoising methods that are used. Each image is of 256x256 dimensions. We have assumed that each image is corrupted with Gaussian noise with three different known variances σ =10, 20 and 30. The image corrupted by noise is decomposed into three levels of wavelet transform by using Daubechies-8 Wavelet. Each sub-band is then applied by the proposed denoising scheme.

TABLE I. PSNR Values of Denoised Images Using Different Methods

Image	σ_n	Wavelet methods using Bayesian Approach [18]	Curvelet Transform using Bayesian Approach [18]	Empirical Bayesian Approach Using HMT [15]	Proposed
Lenna	10	27.58	31.93	30.51	**32.50**
	20	26.02	28.49	28.56	**28.74**
	30	22.03	**26.70**	**26.70**	26.63
Camera Man	10	24.63	30.09	30.42	**30.89**
	20	24.86	26.67	25.32	**27.94**
	30	22.72	24.80	25.42	**25.51**
Peppers	10	29.00	29.93	**32.61**	32.33
	20	22.39	26.46	27.11	**28.85**
	30	18.78	24.67	25.33	**26.60**
Barbara	10	28.79	29.57	-	**31.89**
	20	24.14	26.41	-	**27.88**
	30	22.46	24.69	-	**25.79**

Standard images include Lena, Camera Man, Peppers and Barbara. These images are tested using the proposed method comparable to other techniques including Wavelet Transform using Bayesian Network Approach [18], Curvelet Transform using Bayesian Network Approach [18] and Empirical Bayesian Approach using HMT [15]. Also, some local images are also used that are tested based on the PSNR of the noisy image compared with the denoised one.

It can be noticed from the table I that our proposed scheme gives a better performance in terms of PSNR values. We have compared our method in terms of PSNR at different noise variances. Our technique shows comparable results to that of other techniques. It can be noticed that Empirical Bayesian approach performs well when used in HMT domain. Comparable results are found in proposed work when the HMT is initialized by EM algorithm. Also, the standard image 'Lenna' is tested for visual quality. Using our proposed method, it has been shown in fig. 3 that image quality is comparable to other techniques with less blurring.

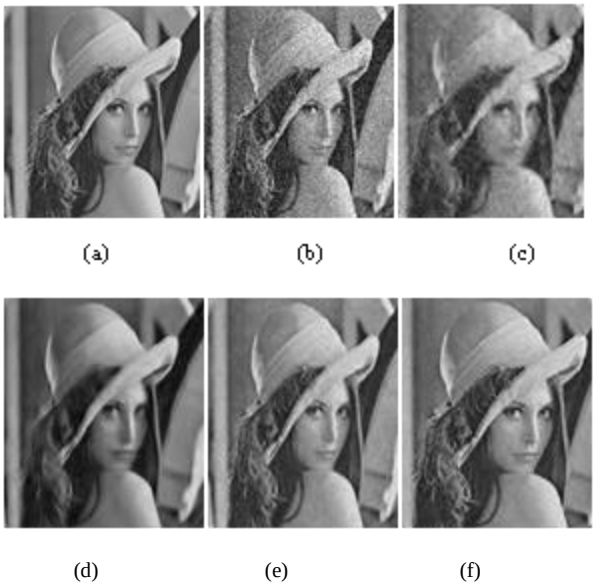
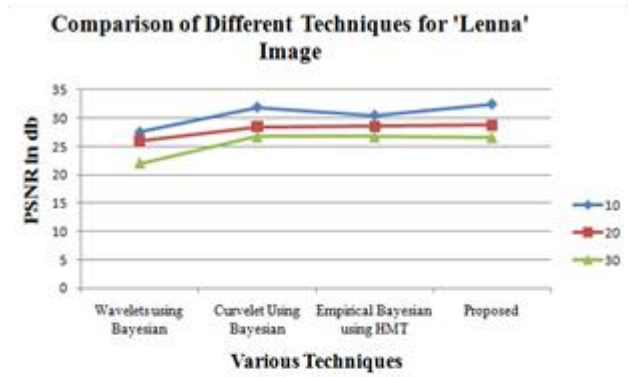
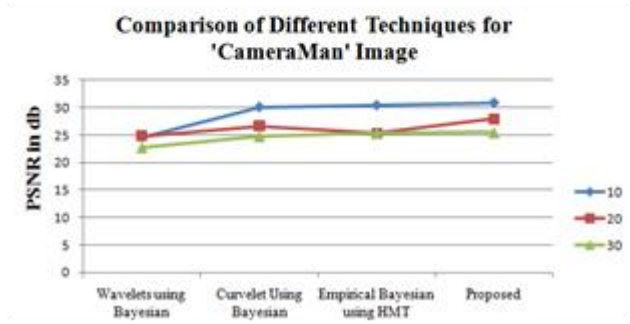


Fig. 3. Images of denoising experiment corresponding to first row of table I. (a) Original 256x256 'Lena' image (b) Noisy Image with σ = 20 , PSNR= 22.41 db. Image Denoised by (c) Wavelet Transform using Bayesian Approach [18], PSNR=26.02db (d) Curvelet Transform using Bayesian Approach [18], PSNR= 28.49db (e) Empirical Bayesian Approach using HMT [15], PSNR= 28.56db (f) Proposed Method, PSNR= 28.74db



(a)



(b)

Fig. 4. Comparison of various denoising methods (a) 256x256 'Lenna' image (b) 256x256 'CameraMan' image

TABLE II. MEAN OPINION SCORE CLASSIFICATION

5	Excellent Quality
4	Good Quality
3	Fair Quality
2	Poor Quality
1	Bad Quality

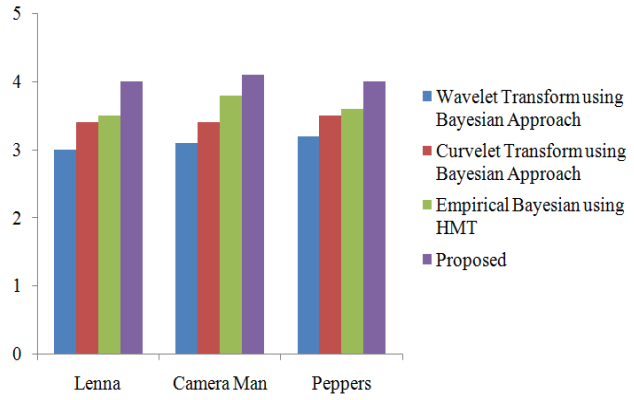
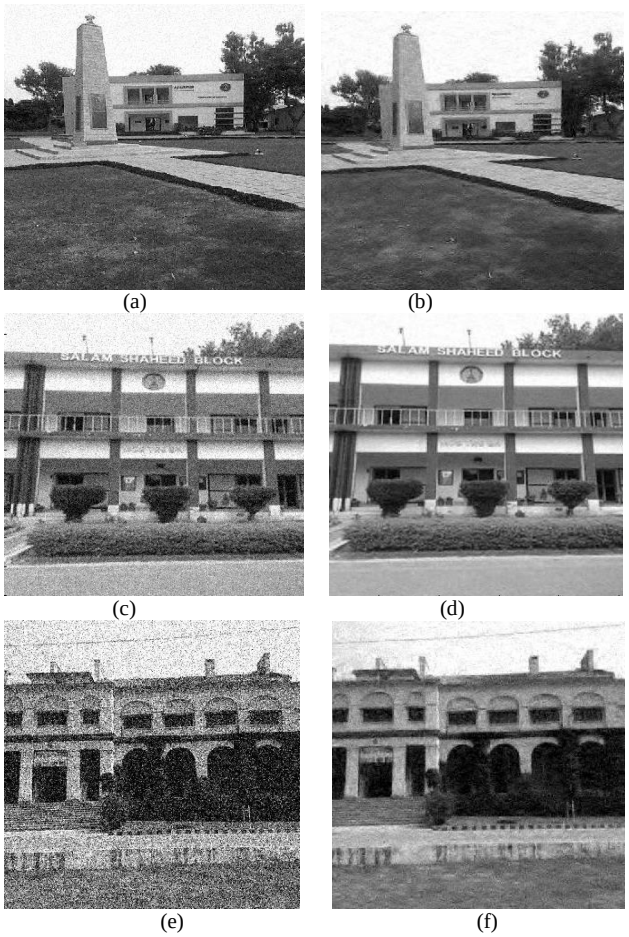


Fig. 5. Graphical Representation of Mean Opinion Score for Test Images

The different denoising techniques have been compared graphically as well. Fig.4, shows the graphical representation of different techniques at 3 different noise variances.

From many years MOS (Mean Opinion Score) method has been used for subjective analysis of an image, video or voice quality. This technique refers to the averaged value of the opinions taken from the users [19]. This method scores the quality of the image from 1 (worse) to 5 (excellent). Table II shows different classes of MOS ranging from 1 to 5.

We have performed MOS analysis on three different images that are tested using different techniques in order to have an idea about visual quality of the images. Fig. 5, shows the results taken from a group of people and then averaged over the total number of observations. The graph shows that visual quality of images using our proposed technique has the comparable score to other techniques.



(a) (b) (c) (d) (e) (f)

Fig. 6. Images of denoising experiment for some local images. (a) Noisy image with $\sigma = 10$, PSNR= 28.12db (b) Denoised Image, PSNR= 30.44 db. (c)Noisy Image with $\sigma = 20$, PSNR= 22.14db (d)Denoised Image, PSNR= 29.53db (e) Noisy image with $\sigma = 30$, PSNR= 18.62db (f) Denoised Image, PSNR= 25.00db

The denoising algorithm proposed is also applied on several images taken from the institute as well. These images are tested at different noise variances and noisy image is compared with the denoised image in terms of PSNR. Fig.6, shows the results taken after applying the algorithm.

IV. CONCLUSION

We have proposed HMM based image denoising method in Time-Frequency analysis domain. It has been shown that the proposed algorithm outperforms the existing few techniques based on PSNR values. The proposed framework concisely models the non-Gaussian statistics of the individual wavelet

coefficients. The EM algorithm used for training the HMT model captures the statistical dependencies between the coefficients that are tested using some standard and local images comparable to other techniques. This technique gives comparable results as compared to HMT that uses Bayesian approach to model the coefficients. Our method shows improvement of the denoised images as compared to other techniques based on both the PSNR values and in terms of visual quality of the images.

ACKNOWLEDGMENT

This research work has been facilitated by the Image Processing cell at Military College of Signals, National University of Sciences and Technology, Islamabad.

REFERENCES

[1] R.C. Gonzalez and R.E Woods: 'Digital Image Processing', Prentice Hall, Upper Saddle River, N.J., Second edition, 2002

[2] M. S. Crouse, R. D. Nowak and R. G. Baraniuk: 'Wavelet-based statistical signal processing using hidden Markov models', IEEE Transaction on Signal Process., vol.46, no.4, pp. 886 -902, 1998.

[3] G. Y. Chen and T. D. Bui: 'Multi1wavelet Denoising using Neighboring Coefficients', IEEE Signal Process. Lett., vol. 10, pp. 211-214, 2003.

[4] Jyri J.Kivinen, Erric B. Sudderth, Micheal I. Jordan, "Image Denoising with nonparametric Hidden Markov Trees", 2007

[5] D. Cho and T. D. Bui: 'Multivariate statistical modeling for image denoising using wavelet transforms', Signal Processing: Image Communication, vol. 20, pp. 77–89, 2005.

[6] Minghui Yang, Zhiyun Xiao, Silong Peng, "A wavelet Domain Hidden Markov Tree Model with Localized Paremeters for image denoising", 2006

[7] D. Cho, T. D. Bui and G. Y. Chen: 'Image denoising based on wavelet shrinkage using neighbour and level dependency', International Journal of Wavelets, Multiresolution and Information Processing, vol. 7, no. 3, pp. 299-311, 2009.

[8] G. Y. Chen, W. P. Zhu and W. F. Xie: 'Wavelet-based image denoising using three scales of dependency', IET Image Processing, vol. 6, no. 6, pp. 756-760, 2012.

[9] G. Y. Chen, T. D. Bui and A. Krzyzak: 'Image Denoising using neighboring wavelet coefficients', Integrated Computer-Aided Engineering, vol. 12, no. 1, pp. 99-107, 2005.

[10] J. Ho and W. L. Hwang, "Wavelet Bayesian network image denoising," *IEEE Transaction on Image Processing*, vol. 22, no. 4, pp. 1277-1290, 2013

[11] Guoliang Fan and Xiang-Gen Xia: 'Wavelet-based Texture Analysis and Synthesis Using Hidden Markov Models', IEEE Transactions on Circuits and Systems-I, Fundamental Theory and Applications, vol. 50, no. 1, January 2003

[12] Gilbert Strang and Truong Nguyen: 'Wavelets and Filter Banks', Wellesley-Cambridge Press, 1997

[13] R. D. Nowak, 'Multiscale hidden Markov models for Bayesian image analysis', in Bayesian Inference in Wavelet Based Models, P. Müller and B. Vidakovic, Eds. New York: Springer Verlag, 1999, pp. 243–266.

[14] H.Choi and R. Baraniuk: 'Multiscale image segmentation using wavelet-domain hidden Markov models', IEEE Trans. Image Processing, vol. 10, pp. 1309–1321, Sept. 2001.

[15] J. K. Romberg, H. Choi, and R. G. Baraniuk: 'Bayesian tree-structured image modeling using wavelet-domain hidden Markov models', IEEE Trans. Image Processing, vol. 10, pp. 1056–1068, July 2001.

[16] H. Chipman, E. Kolaczyk, and R. McCulloch: 'Adaptive Bayesian wavelet shrinkage', J. Amer. Stat. Assoc., vol. 440, no. 92, pp. 1413–1421, Dec. 1997

[17] M. Amini, M.O. Ahmad and M.N.S. Swamy, "Image denoising in wavelet domain using the Vector-based hidden Markov model," IEEE12th Inter. New Circuits and Systems Conf. (NEWCAS), pp. 29-32, 2014.

[18] Pallavi Sharma, R.C. Jain, Rashmi Nagvani, "An efficient Curvelet Bayesian Network approach for Image Denoising," IEEE International Conference on Advances in Engineering & Technology, 2014

[19] Anna Geomi George, A. Kethsy Prabavathy, "A Survey on Different Approaches used in Image Quality Assesment", International Journal of Emerging technology and Advanced Engineering, vol.3, Issue 2, February 2013.

Connected Dominating Set based Optimized Routing Protocol for Wireless Sensor Networks

Hamza Faheem*, Naveed Ilyas†, Siraj ul Muneer‡, Sadaf Tanvir§
*†‡Dept. of Computer Science
COMSATS Institute of Information Technology, Park Road Chak Shahzad, Islamabad,Pakistan
§Dept. of Computing and Technology, IQRA University, Plot no. 5, H-9, Islamabad,Pakistan

***Abstract*—Wireless Sensor Networks(WSNs) have problem of energy starvation in their operations. This constraint demands that the topology of communicating nodes should be limited. One of the ways of restraining the communicating nodes is by creating a Connected Dominating Set(CDS) of nodes out of them. In this paper, an Optimized Region Based Efficient Data(AORED) routing protocol for WSNs has been proposed. CDS has been employed in AORED to create a virtual backbone of communicating nodes in the network. The empirical study involving extensive simulations show that the proposed routing protocol outperforms the legacy DEEC and SEP protocols. AORED has increased number of transmission rounds, increased number of clusterheads and reduced number of packets sent to the basestation as compared to DEEC and SEP protocols.**

***Keywords*—Connected Dominating Set; Wireless Sensor Net-works; Energy Efficiency**

I. Introduction

Wireless sensor networks (WSNs) have grabbed attention of researchers in recent years due to their wide range of applications and potential use. A WSN comprises of independent sensor nodes that sense and transfer physical information to the sink. Cheap and small sized sensor nodes cannot be equipped with a large battery source. As these sensors have limited energy, it restricts the sensors to use limited memory, limited transmission power and perform limited computations to increase the lifetime of sensor. Network and data link layers in a sensor node play a vital role in the WSN communication. The energy problem is usually solved by the sensor node by changing its state. Generally, there are three states of sensor nodes i.e. active state, idle state, and sleeping state. The main purpose of active state is to transmit and receive data packets. Active state uses the maximum amount of resources available to the nodes which results in energy consumption. During the idle state, the only task done by the nodes is of sensing without receiving or transferring the data. While in the sleep mode, maximum energy is saved by turning of its radio off by the nodes. Sleep mode consumes the minimum amount of energy as compared to active and idle mode. Another way to conserve nodes' energy is to use energy efficient routing protocols and use limited transmit power if possible that reduces transmission range.

The issue of energy limitation in WSNs has been addressed by using the concept of Connected Dominating Set(CDS) as well. The CDS limits the number of communicating nodes hence reducing the energy consumption. In this paper, a connected dominating set based optimized routing protocol for WSNs has been proposed. Rest of the paper is organized as follows: Related work is presented in section II followed by proposed technique in section III. Section IV presents network model and problem formulation followed by energy consumption in section V and performance evaluation in section VI. Conclusions and future work are presented in the end.

II. Related Work

Routing protocols in WSNs are divided in three major categories; the location based routing, flat routing and hierarchical routing protocols. In case of location based routing protocols, the information of location of the node in WSNs is used to find the distance between nodes. It helps to select the next relay node. The location of node assists the protocol to send data to the specific location, which prevents energy consumption of the whole network. In flat routing protocols, all the nodes play the same role in the network. The BS sends the query to a specific node and that specific node responses to the query. All the nodes send data to neighboring nodes who send it to the BS. The major disadvantage of this routing protocol is that each node sends its data and forwards many other nodes' data to BS. This mechanism drains the energy of the whole network very quickly. In hierarchical routing protocols, group of nodes join together to make a cluster. Among them, there is a cluster head (CH). CH is responsible for receiving data from the nodes. It aggregates and forwards the data to BS. The main drawback of this scheme is CHs high energy consumption due to its additional workload. Random selection of CHs and their rotation from time to time overcomes this issue. However, the formation and selection of CH requires additional energy.

Low Energy Adaptive Clustering Hierarchy (LEACH) [1] is one of the most famous hierarchy based routing protocol in WSNs. Improvements have been proposed on LEACH. One of which is Stable Election Protocol(SEP) [2]. SEP is based on weighted election probabilities of each node to become cluster head according to the remaining energy in each node. Another improvement on LEACH is the Distributed Energy Efficient Clustering(DEEC) routing protocol [3] which selects the CH for the nodes with different levels of the energy. The CH selection is based on residual energy of the node over the average energy of the network. Therefore, the node which has high initial and residual energy has more chances to become

a CH for the particular round than the node which has lower energy.

Apart from these two famous approaches, we see a lot of work done in the recent past to improve the energy efficiency of LEACH. In this regard, [4] presents a new energy-efficient cluster-based routing protocol, which adopts a centralized clustering approach to select cluster headers by generating a representative path. To support reliable data communication, they present a multihop routing protocol that allows both intra- and intercluster communications. In [5], the authors formulate the shortest path routing and the least energy cost routing in wireless sensor networks as L1-norm and L2-norm optimization problems to maintain the maximum network life time. In [6], the authors have proposed to measure the connectivity of sensor nodes and use this parameter for selection of cluster head for forwarding data to BS. In [7], they have used particle swarm optimization with LEACH to increase network life time. In [8], the base station finds the highest energy node among the cluster and mark it as a cluster head for the current time to improve energy efficiency of the network. In [9], the authors propose regional energy aware clustering with isolated nodes for wireless sensor networks. The cluster-heads are selected by the calculated weights based on the residual energy of each sensor and the regional average energy of all sensors in each cluster. In [10], they propose optimum number of cluster heads based on minimizing the dissipated energy in all phases of communication.In [11], The CH selection formula used in [1] is used. However, during re-clustering, a threshold value will be used to decide whether the CH will be replaced or not. Another improvement of leach is proposed in [12]. EELP is proposed for critical applications. Sensor nodes are manually deployed in each room on the floors of the multistorey apartment block. The nodes deployed in the rooms on each floor are assumed as a separate cluster and the node with the highest energy is selected as the CH to decrease the probability of the selection of a node with low energy and to balance the total energy load distribution of the network. In [13], only nodes with maximum residual energy and minimum energy consumption can become cluster heads since each nodes residual energy as well as average energy consumption is considered for the selection of cluster heads. In [14], the first level CHs are elected as in LEACH protocol. After the election of first level CHs, the second level CHs are elected hence introducing two-level hierarchy of CH nodes. The data undergoes multiple hops among CHs thereby increasing network lifetime. In [15], LEACH is improved by introducing master head and shortest path algorithm. Sensor nodes send data to the BS using MIMO. In [16],H-LEACH, the nodes with energy less than to that of the minimum energy required for transmitting and receiving signals is made to die as it lacks energy to do it. Minimum threshold is subtracted from the energy of the node in every round as that much of energy is consumed. Total number of alive nodes are calculated for every round so as to have a track on the life time of the network. In [17], LEACH-MAC, attempts to control the randomness present in LEACHs clustering algorithm by using MAC layer function. This approach makes the cluster head count stable. In [18]LEATCH, a two level hierarchical approach has been proposed to organize a sensor network into a set of clusters, every cluster divided into small clusters that are called Mini Clusters. As the way the clusters are organized, for each mini cluster, we define a Mini Cluster-Head (MCH). Every MCH communicates with the cluster-head directly, it aggregates its mini-cluster information and passes it on to the base station. In [19]P-LEACH, it combines the features of LEACH and PEGASIS [20] to improve energy efficiency in routing. In [21], each node broadcasts a Hello message which contains its identity and energy status with a predefined transmit power. With the received power of Hello messages and the prior knowledge of the transmit power, nodes can estimate the distance among them. Based on distance profiles, nodes create a list of N dominant neighbors. After this, the same steps are performed as that of LEACH. [22] present a comprehensive overview of different approaches under structure-free and structured wireless sensor networks for data collection and aggregation, clustering and routing along with their key issues. In [23], the authors propose to keep the cluster head count optimal using a cross layer approach since this count directly affects the energy efficiency of the network. In [24], they choose clusterheads based on their geographical locations to improve the overall network performance.

The limited power constraint in WSNs demand that the topology of communicating nodes should be limited. This need has urged the researchers to design new algorithms for controlling topology of the network. Connected Dominating Set (CDS) based topology control has received attention to reduce redundant and unnecessary communication overhead. Having such a CDS restricts the main communication tasks to the dominators only. In recent past, many efforts regarding the usage of CDS in wireless sensor networks can be seen. In [25], they consider how to construct a CDS in WSNs. In [26], they summarize the various CDS constructing algorithms both centralized and distributed and compares them. In [27], the authors propose a new degree-based greedy approximation algorithm to construct the CDS and then reduce its size by excluding some of the CDS nodes cleverly without any loss in coverage or connectivity. In [28], they have considered the problem of minimizing the number of CDS vertices that belong to a subset $v \subset V$. In [29], the authors present a novel algorithm based on the Induced Tree of Crossed Cube to reduced the size of CDS. In [30], they have performed a comparative study of major works relating to CDS construction emphasizing on the type of algorithm, technique employed, performance metric used and the outcome achieved.

III. Proposed Technique

In this paper, An Optimized Region Based Efficient Data Routing (AORED) protocol has been proposed in which the whole network field is divided into two regions to optimize the transmission power of sensor nodes. This partition is based on the euclidean distance to BS. AORED employs CHs to send data to base station. CHs in both regions transmit the aggregated data with the same transmission power to BS. CHs that are faraway from the BS, first send data to routing nodes which then relay the data to BS.

A. Procedure

The proposed protocol AORED is based on rounds; each round consist of two phases: setup phase and steady phase. The proposed protocol introduces three additional concepts: Logical formation of two groups, selection of CHs among the

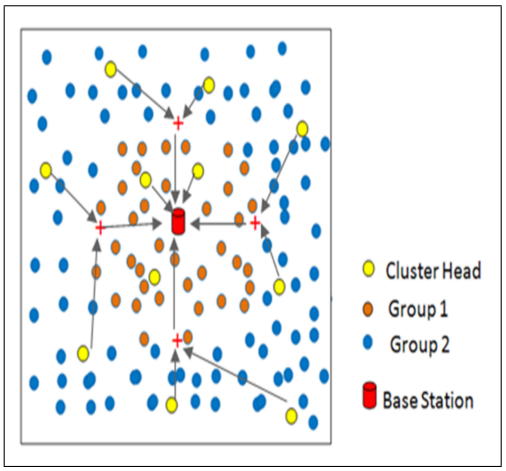


Fig. 1: AORED Communication Scenario

same group and installation of four routing nodes in the center of sensor field (distance measured from the BS). Additional concepts and phases in AORED are discussed in detail below.

B. Groups Formation

The whole network field is divided into fixed logical groups. These groups are homogenous in nature; each group may have different number of sensor nodes. BS is placed in the center of sensor field. The formation of groups requires the distance to BS of each node. If the distance of node to BS is less than or equal to 30% of the network size $100m \times 100m$, these nodes will be the part of group 1 shown in orange color in figure 1. All other nodes that lie at a distance greater than 30% of distance to BS belong to group 2 in blue color.

C. Routing Nodes Installation

The routing nodes have the same capabilities as normal nodes but they are used for a specific purpose. The four routing nodes are placed in the center of the sensor field along the x and y axes. + sign shows routing nodes in figure 1. Routing nodes neither take part in the CH formation nor in any other sensor field operations. They are installed to route the data of group 2 CHs. The group 1 CHs directly communicate to BS with the normal transmission power. In the same way, group 2 CHs also forward the data to routing nodes with the normal transmission power. The routing nodes aggregate the data received from the group 2 CHs and forward it to BS. Before forwarding the data to routing nodes, group 2 CHs calculate the distance to the routing nodes and forward it to the nearest routing node.

In figure 1, the total area is $100m \times 100m^2$ and the BS is in the centre at position of (50,50). Routing nodes are installed at the position of (25,50), (50,25), (75,50) and at (50,75).

D. Setup Phase

In the setup phase, it uses the same mechanism of CH selection as described in [1]

Grouping Phase:
Node $i \in$ N broadcasts ranging probe to BS
BS unicasts reply to i
i calculates distance $di \rightarrow$BS
if $di < 30$ **then**
if $i_{Energy} > 0$ **then**
i joins Group 1
else
i joins Group 2
endif
endif

Fig. 2: Algorithm 1

$$T(n) = \begin{cases} \dfrac{p}{1-(p*rmod(1/p))} & if\ n \in G \end{cases} \quad (1)$$

$$T(n) = 0 \quad if n \notin G \quad (2)$$

In equation 1, p is the percentage to be a CH, G is a group of nodes that have not served as a CH since $1/p$ rounds so far and r is the present round. After a few rounds, the energy of the nodes will be uneven and nodes who have not taken the chance will become the CHs. AORED protocol provides this function to work with specific group. If the region of the node is group 1, then it can be a member of group 1 CH and if the node lies in group 2 region, then it can join the CH of group 2.

E. Steady Phase

In this phase, the communication starts between the nodes and respective CHs according to TDMA slots allocation. All the nodes only communicate through their CHs. Elected CHs broadcast their status using CSMA/CA protocol. Non-CH nodes select their CHs by comparing the strength of received signals from multiple CHs. After creating clusters, all CHs will create TDMA schedule for their associated members and broadcast it. All those nodes whose time slot is not active, are in sleep mode to conserve energy. The group 1 region CHs in this phase aggregate the data and forward it to the BS. On the other hand, the group 2 region CHs aggregate the data and forward it to the routing nodes. After receiving data from many CHs, the routing nodes aggregate the data and forward it to the BS as shown in figure 1.

The algorithm for grouping and steady phase of AORED is given in figure 2 and 3.

IV. Network Model and Problem Formulation

For problem formulation, the network area is divided into two groups: group 1 and group 2. An undirected graph G = (V, E), such that V is the set of vertices and E is the edges of graph G. In this scenario, there are two types of vertices V_{CDS} and V_{nonCDS}. Similarly the edges E are also divided into E_{CDS}

Steady Phase:
if $i \leftarrow$N **then**
TransToCH(i,Datapack)
end if
if $i \leftarrow$CH **then**
CH(i) estimates distance to routing nodes(RN)
CH(i) calculates distance $D_{CH \rightarrow RN}$
$D_{CH \rightarrow RN} = min(d_{i \rightarrow CH})$ *note down the RN that is at the lease distance away from node*
RCV$_{CH}$*(i,Datapack)CH receives data packet from node* i
Aggregate$_{CH}(i, Datapack)$
TransToRoutingNodeCH(i,Datapack)
end if
if $i \leftarrow$Routingnode(RN)**then**
RCV$_{RN}$(i,Datapack)*RN receives data packet from node* i
AggregateRN (i,Datapack)
TransToBaseStation (i,Datapack)
end if

Fig. 3: Algorithm 2

and E_{nonCDS} such that, E = $E_{CDS} \bigcup E_{nonCDS}$. Vertices V of graph G can be defined as $V = V_{CDS} \bigcup V_{nonCDS}$. The CDS nodes are selected from the group 1 region of 30 meters. A Dominating Set (DS) of graph *G* is a subset *D* of *V* such that every vertex in *V / D* is adjacent to at least one member vertex of *D*. A CDS is defined as a subset *D* of *V* such that any node in *D* can access any other node in *D* by a path that lies entirely within *D*, such that *D* induces a connected sub graph within *G*.

Let n number of nodes be randomly deployed in a two dimensional area. The nodes are categorized into three types: Normal Nodes *(NN)*, Cluster Head *(CH)* and Routing Node *(RN)*. *(NNs)* sense the data and transmit to *CHs*. *CHs* forward the data packets to *RNs*. These *RNs* forward the data packets to BS lying at the centre of network field.
According to mathematical models presented in literature, the network's throughput can be maximized using the equation 3.

Maximize:

$$d_{total} = \sum_{i=1}^{NN} u_t^i + u_r^i + \int_0^{2\pi} \int_0^r p(\pi r^2) r dr d\theta \qquad (3)$$

Subject to:

$$\sum_{i=1}^{n} u_t^i + u_r^i \leq E_{total} \quad \forall i \in i \qquad (4a)$$

$$\sum_{c=1}^{CH} \sum_{i=1}^{NN} f_{ci} \leq F \quad \forall c, i \in n \qquad (4b)$$

$$f_{ci} \leq C_{ci} \quad \forall c, i \in n \qquad (4c)$$

$$\bar{e}, d_{ci} \geq 0 \qquad (4d)$$

Eq. 3 describes that maximum number of nodes should live for far longer duration in order to increase the throughput. Eq. 4a depicts that energy spent by node i to transmit and receive u bits is upper bounded by total energy given to network. Eq. 4b describes that data flow between *NN* and *CHs* is upper bounded by total flow of the network. f_{ci} is the flow from NN to CH. Eq. 4c elaborates the relation between flow and total capacity of a particular link. C_{ci} represent the total capacity of link.

A. Criterion for CDS Construction

A Connected Dominating Set(CDS) is developed by using extended localized algorithm presented by Dai and Wu in [31]. After establishment of CDS, the length between all the CDS nodes is calculated. The total distance of all the CDS nodes is calculated by using euclidean distance formula between nodes *i* and *j*. CDS formation includes following steps:
Criteria 1: Nodes with the highest degree are identified in the inner region. After that, consider those nodes which have highest degree and then second highest degree. The highest degree nodes are called dominator nodes. The nodes which are adjacent to highest degree nodes are called dominatee nodes.
Criteria 2: Dominatee nodes which are not further adjacent to any other node are called leaf node. The leaf nodes are never included in CDS.
Criteria 3: Dominatee nodes other than leaf nodes are converted to dominator nodes if that dominatee node has only leaf neighbors.
Criteria 4: Node which is a dominatee of two dominator nodes is also converted to dominator node.
Definition 1: A set of nodes $V_{CDS} \subset V$ such that every node $\hat{v}$ belongs to $V - D$, there should be at least one node $\hat{u}$ in V_{CDS} that dominates $\hat{v}$. Furthermore, V_{CDS} are connected to each other.
Definition 2: A CDS,$V_{CDS} \subset V$ of G = (V, E),such that CDS nodes act as RNs and non-CDS nodes act as normal nodes which forward data to RNs. The set of RNs collectively form a random path within CDS through which data is routed to the static sink that exists at the centre of network field.

Definition 3: A CDS, $V_{CDS} \subset V$, such that the random path exists within CDS nodes is converted to a circular path formed by CDS nodes. This circular internal area is optimized to achieve:

$$Max d_{total} = \sum_{i=1}^{NN} u_t^i + u_r^i + \int_0^{2\pi} \int_0^r p(\pi r^2) r dr d\theta \qquad (5)$$

Consider the example in which nodes are randomly deployed in network field. As the network has two regions, the group 1 region is optimized to achieve the maximum throughput. For this purpose, consider 15 nodes inside group 1 region and then apply the CDS formation rule as discussed above. By searching the first and second highest degree nodes inside the circular region, one finds out that node n_7 has highest degree and node n_{13} has second highest degree. These two nodes are called dominator and thus included in CDS. After completion of first step, the second step is to sort out the dominatee nodes which are adjacent to dominator nodes. The dominatee nodes of n_7 include $(n_4, n_2, n_5, n_6, n_1, n_9)$. The leaf nodes n_5, n_6 are not included in the CDS. Nodes n_4 and n_2 have leaf neighbors, so

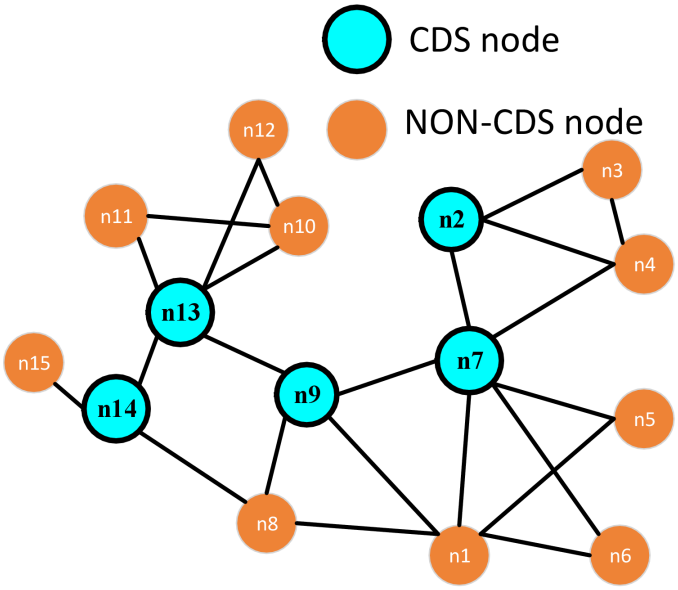


Fig. 4: Construction of CDS

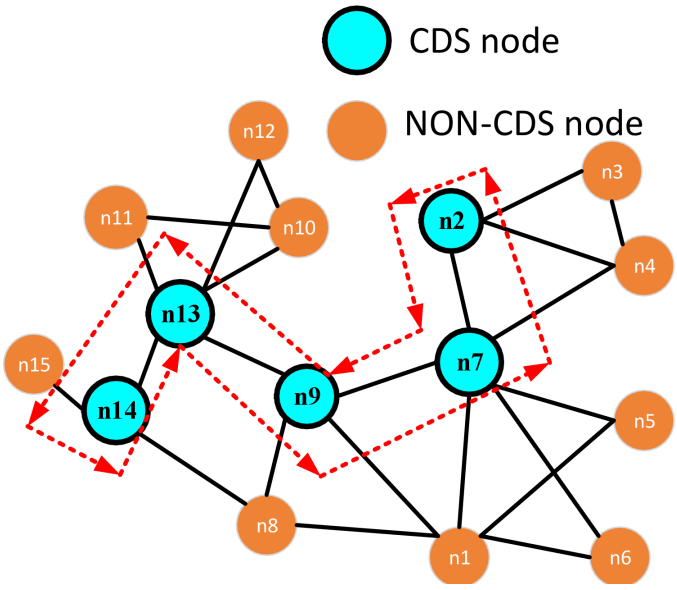


Fig. 5: Random Path formed by CDS nodes

any one of them is included in CDS. Thus, node n_4 is included in CDS. The dominatee nodes n_1 and n_9 have further non-leaf neighbors so these two nodes are not included in CDS.

Consider the second highest degree node n_{13}, and according to criteria 1, the node n_{13} is included in CDS. The dominatee nodes of n_{13} include $(n_{10}, n_{12}, n_{11}, n_{14}, n_9)$. According to criterion 2, nodes n_{10} and n_{12} are leaf nodes and hence these two nodes are not included in CDS. The node n_{14} is included in CDS because it has leaf neighbor that is node n_{15}. Moreover, according to criteria 4, the node which is dominatee of two dominator nodes is also included in CDS, thus node n_9 is included in CDS. After implementation of all the above mentioned rules on the group 1 region of network, we obtain the CDS which includes nodes $(n_2, n_7, n_9, n_{13}, n_{14})$ as shown in figure 4. The Path length (PL) of all the CDS nodes can be calculated as shown in eq 6. Figure 5 shows the random path constructed within CDS.

$$P_L = \{e(n_2, n_7), e(n_7, n_9), e(n_9, n_{13}), e(n_{13}, n_{14}) \\ e(n_{14}, n_{13}), e(n_{13}, n_9), e(n_9, n_7), e(n_7, n_2)\} \quad (6)$$

The length of all the CDS nodes can be computed by adding the length of all links of CDS as follows:

$$P_L = \sum_{\forall e(i,j) \in CDS} length\ (i,j) \quad (7)$$

Where *length(i, j)* $\in$ *CDS* is the distance between node *i and j*. $length(i, j)$ is computed by using Euclidean distance formula:

$$d(i,j) = \sqrt{(x_i - x_j)^2 + (y_i - y_j)^2} \quad (8)$$

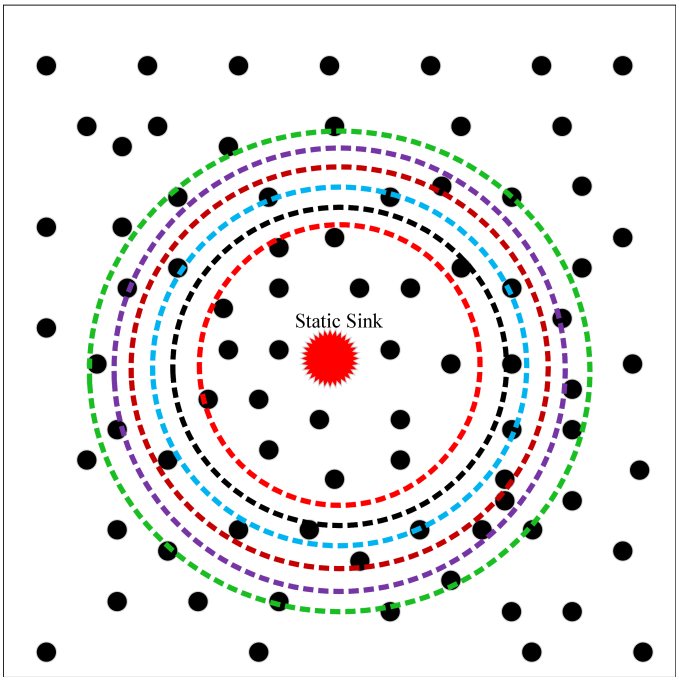


Fig. 6: Conversion of CDS into Circular Shape

In other words:

$$length(i, j) = d\ (i, j) + d(j, i) \quad (9)$$

In this way, total path length of all the CDS nodes is computed by using following equation:

TABLE I: Different values of α

α	r	r = r + α
0	30m	30m
0.2	30m	36m
0.4	30m	42m
0.6	30m	48m
0.8	30m	54m
1	30m	60m

$$P_L = 2 \times \sum_{\forall e(i,j) \in CDS} \textit{length (i,j)} \tag{10}$$

B. Optimized Group 1 Region

Irregular trajectory formed by CDS nodes can be changed to a circular internal area as follows:

Since area of a circle is:

$$A = \pi r^2 \tag{11}$$

Where *r* is a radius of circle. To compute *r*, circumference of a circle is C= $2\pi r$.

Where:

$$r = \frac{C}{2\pi} \tag{12}$$

$$C = P_L = 2 \times (\sum_{\forall e(i,j) \in CDS} \textit{d(i,j)}) \tag{13}$$

$$r = \frac{2 * (\sum_{\forall e(i,j) \in CDS} \textit{d(i,j)})}{2\pi} \tag{14}$$

$$r = \frac{\sum_{\forall e(i,j) \in CDS} \textit{d(i,j)}}{\pi} \tag{15}$$

In order to find the optimized value of circular path, equation 15 will be used

$$r = r + r\alpha \qquad 0 < \alpha, < 1 \tag{16}$$

Let the initial value of r = 30 meter. The value of α will be varied from 0 to 1 as shown in table I. AORED protocol has been simulated for different values of α. Figure 6 shows the different circular paths made by *RNs*. The value of α affects the location of RNs. Moreover, the value of α plays a significant role to increase the stability period as well as throughput of network. The value of α is varying as shown in table I. When the value of α is equal to 0, the *RNs* are far from the nodes that exist at the corner of the field. The goal is to find an optimal circular position for *RNs* in order to decrease the transmission distance between CHs in the outer region and *RNs*. When distance between *CHs* and *RNs* decreases, the nodes live for longer duration and hence transmit data packets for longer duration which increases the throughput. By simulating AORED protocol for different values of α, one finds out that at $\alpha = 1$, the stability period as well as network throughput increases. The optimal area of group 1 region can be calculated as follows:

$$A = \frac{(\sum_{\forall e(i,j) \in CDS} \textit{d(i,j)})^2}{\pi} \tag{17}$$

V. Energy Consumption

There are three different types of nodes: normal, routing and CH nodes depending upon their role. Energy consumption of these nodes is given in next subsections.

A. Energy Consumption of a Normal Node (NN)

The energy consumption of NN can be calculated as:

$$E_{NN} = \bar{e}_s + \bar{e}_t(d_{NC}) \tag{18}$$

$$E_{NN} = \sum_{NN}^{n} \bar{e}_s + \bar{e}_t(d_{NC}) \tag{19}$$

E_{NN} in equation 18 represents the energy consumption of NN to sense and transmit their own data. $\bar{e}_s$ and $\bar{e}_t$ are the energies required to sense and transmit the data. The distance between NN and CH is represented by d_{NC}. If all the normal nodes are considered in the network, then energy consumption can be represented as in equation 19. Similarly, the energy consumption of NN to forward data to BS can be computed as:

$$E_{NN} = \bar{e}_s + \bar{e}_t(d_{NB}) \tag{20}$$

$$E_{NN} = \sum_{NN}^{n} \bar{e}_s + \bar{e}_t(d_{NB}) \tag{21}$$

where d_{NB} in equations 20 and 21 represents the distance between *NN* and *BS*.

B. Energy Consumption of CH

For CHs, it is assumed that they share equal load from NNs within the cluster. First consider the energy consumption rate for nodes in group 2. Equation 22 depicts the energy consumption of CHs in group 2. However, equation 23 shows the energy consumption of CHs in the group 1.

$$E_{outer}, CH = \bar{e}_s + \bar{e}_t(d_{CR}) + \frac{pn_{outer}}{m_{outer}}[\bar{e}_r + \bar{e}_t(d_{CR})] \tag{22}$$

$$E_{inner}, CH = \bar{e}_s + \bar{e}_t(d_{CB}) + \frac{pn_{inner}}{m_{inner}}[\bar{e}_r + \bar{e}_t(d_{CB})] \tag{23}$$

TABLE II: Simulation Paramters

Parameters	Values
Network size	$100m^2$
Packet Size	1 Byte
Initial Energy	500mJ
Data aggregation Energy Cost	50 pj/bit
Number of Nodes	100
Node Density	0.01
Transmit Electronics	50 nJ/bit
Receiver Electronics	50 nJ/bit
Transmit Amplifier E_{amp}	100 pJ/bit/m^2
Simulation Rounds	3000s

$\bar{e}_r$ is the energy required to receive the data, d_{CR} is the average distance between CH and RN andd_{CB} is the average distance between *CH* and *BS*.

C. Energy Consumption of Routing Nodes (RN)

For RNs, it is assumed that they share equal load from CHs of group 2. Equation 24 depicts the energy consumption of RNs.

$$E_{1,RN} = \bar{e}_s + \bar{e}_t(d_{RB}) + \frac{\sum_{k=2}^{K} pn_k}{m_i nner}[\bar{e}_r + \bar{e}_t(d_{RB})] \quad (24)$$

Where d_{RB} is the average distance between *RN* and *BS*.

VI. Performance Evaluation

Extensive simulations have been conducted in MATLAB to compare AORED protocol with DEEC and SEP. AORED outperforms both DEEC and SEP in energy consumption, network lifetime and network stability. The details of simulation parameters and results are discussed in next subsections.

A. Simulation Parameters

All nodes are deployed using the random uniform distribution within the field of $100m \times 100m^2$. The BS is centrally located at the location of (50, 50). Routing nodes are located at (25, 50), (50, 25), (50, 75)and(50, 75). Total number of nodes in the network is 104 including routing nodes.Figure 1depicts randomly deployed network. Each result is an average of 10 simulation runs. Simulation parameters are listed in table II

To evaluate AORED protocol, the following network parameters have been considered:

1) **Stability Period:** observe the network operation until its first node is dead
2) **Network Lifetime:** observe the time period from the start to the death of all the nodes.

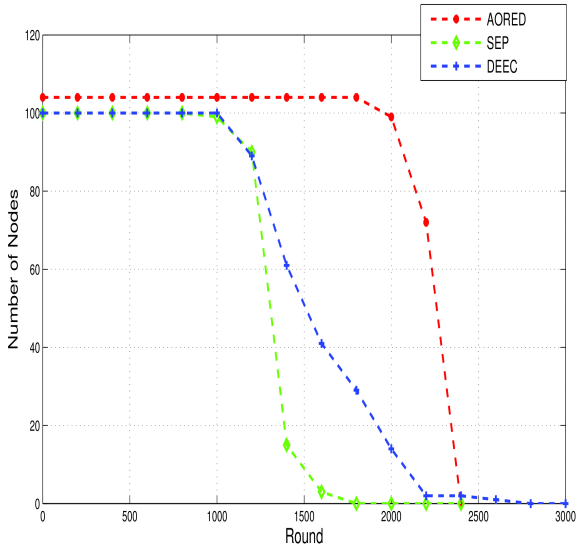


Fig. 7: Number of Alive Nodes

3) **Number of CHs per Round:** observe the number of CHs per round having capability of aggregating the data and sending it to BS.
4) **Number of Alive Nodes:** observe the number of all nodes in WSN that are alive in a particular round.
5) **Variable network size:** observe the behavior of proposed protocol with different network sizes. i.e. $200 \times 200m^2$, $300 \times 300m^2$, $400 \times 400m^2$, and $500 \times 500m^2$ keeping the node density constant i.e. 0.01.

B. Simulation Results

AORED works in rounds as DEEC and SEP. Total rounds used for the experiments are 3000.

Network Lifetime

Figure 7 and 8 show that the AORED has higher network life time as compared to DEEC and SEP. The first node of AORED is dead after 1900+ rounds, whereas in DEEC and SEP, it's 1000+ and 900+ respectively. AORED outperforms the DEEC and SEP in network stability and in network life. Last node of DEEC, SEP and AORED is lifeless at approximately 1674, 1899 and 2371 rounds respectively. AORED has 22% more rounds as compared to DEEC and 24% more rounds than SEP.

$CHSelectionPerRound$

DEEC, SEP and AORED prefer the distributed *CH* selection algorithm. If the algorithm chooses a small number of selected *CHs*, it means each CH forwards more nodes' data. In this way, *CHs* battery diminishes at a fast rate. Selected *CHs* have to perform this additional function of data aggregation and forwarding. If huge number of *CHs* are chosen, it causes overall network energy utilization. Random selection of *CHs* is depicted in figure 9. As there are more rounds in AORED, no. of *CHs* is also increased.

$Packets\ to\ BS$

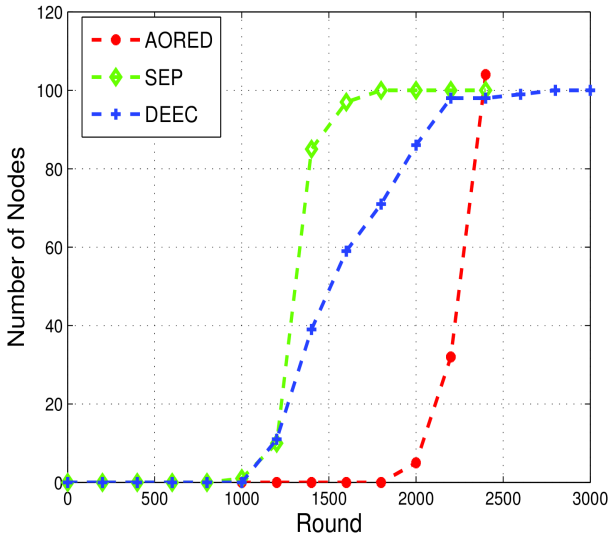


Fig. 8: Number of Dead Nodes

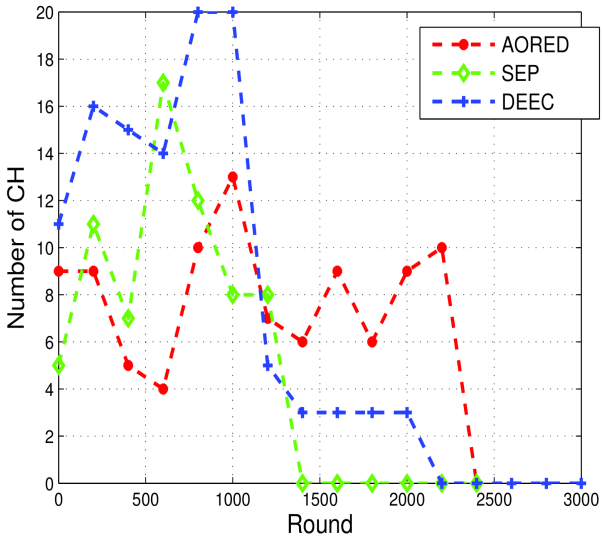


Fig. 9: Increased no. of Clusterheads

In AORED, nodes in group 2 send the data to the Routing(R) nodes. *RNs* after receiving the data from multiple *CHs* aggregate it and send it to *BS*. In this way, almost 50% of packets sent to *BS* are controlled by the *RNs*. Figure 10 shows the impact of packet to *BS* and it can be seen that the no. of packets sent to *BS* in case of AORED have been reduced by 41% as compared to DEEC and 46% as compared to SEP.

AORED with Variable Network Size

AORED has been tested with increasing network sizes. Table III shows the details of various network sizes, no. of deployed nodes and no. of rounds achieved in each network size. In large network sizes, nodes have been grouped based on 30% of distance to *BS* for that particular network size. e.g. in a $200 \times 200m^2$ network, *BS* is located in the center

TABLE III: Increasing Network Sizes

Sr	Region	Nodes Deployed	Rounds
1	100 ∗100	100	2346
2	200 ∗200	400	2363
3	300 ∗300	900	2480
4	400 ∗400	1600	2661
5	500 ∗500	2500	2663

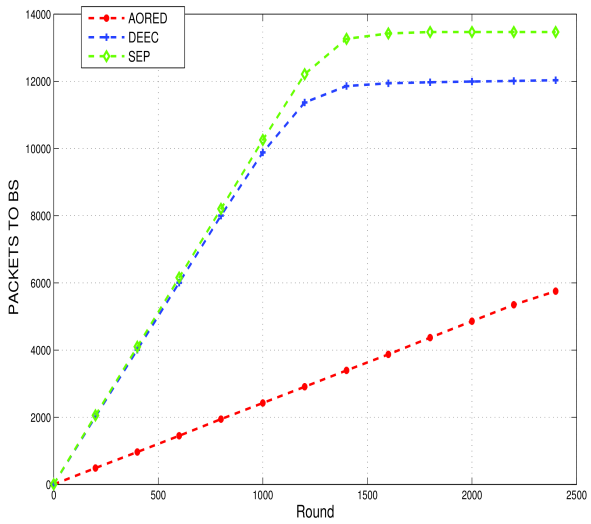


Fig. 10: Packets to BS

of the network at (100,100). Nodes that lie at a distance of $60m(30\%of200m)$ make group 1 and the rest make group 2. Similarly, in a $300 \times 300m^2$ network, *BS* is located in the center of the network at (150,150). Nodes that lie at a distance of $90m(30\%of300m)$ make group 1 and the rest of nodes make group 2. Large network sizes follow the same method of group formation. In each network, 0.04% of nodes are routing nodes e.g. in $200 \times 200m^2$ network, there are 8 routing nodes. Similarly, in a $300 \times 300m^2$ network, there are 12 routing nodes and in $400 \times 400m^2$ network, there are 25 routing nodes. Each result is an average of 10 simulation runs.

Table III shows an increase in the no. of rounds for large network sizes i.e. $300 \times 300m^2$, $400 \times 400m^2$, $500 \times 500m^2$. In a $100 \times 100m^2$ network, nodes who wish to transmit a one bit message at a distance use the following communication model:

$$E_{Tx}(k;d) = E_{elec} * k + E_{fs} * k * d^2 \tag{25}$$

if d $\geq d_o$

$$E_{Tx}(k;d) = E_{elec} * k + E_{amp} * k * d^4 \tag{26}$$

Where E_{Tx} in equation 25, and 26 is the energy spent in transmission, E_{elec} is the energy spent in node's circuits and

E_{amp} is the transmit amplifier energy. Therefore, according to the transmission mechanism mentioned in [1] over which AORED protocol has been based, nodes lying at a distance of less than the threshold of 87m transmit using the equation 25 while the nodes at a distance greater than 87m transmit using the equation 26 [1]. In case of AORED, for small network sizes i.e. $100 \times 100m^2$, $200m^2$, $300 \times 300m^2$, nodes in group 1 and in group 2 both lie at a distance of less than the threshold value i.e. 87m hence, they transmit with the same transmission power and die at almost the same time. However, in case of $400 \times 400m^2$ and $500 \times 500m^2$ networks, nodes in group 1 transmit using equation 25 and nodes in group 2 transmit using equation 26 i.e. with high power as compared to the nodes in group 1. Therefore, energy of nodes in group 2 drains out quickly as compared to the nodes in group 1 and that extends the no. of rounds.

VII. Conclusions and Future Work

In this paper, an optimized region based routing protocol has been proposed which focuses on improving the routing process by reducing the communication overhead in WSN. Communication overhead has been reduced by employing CDS that creates a virtual backbone of communicating nodes in the netwok. This paper provides a mathematical model for maximizing network throughput, stability and life. The model is then verified by extensive simuations. Simulation results show that the proposed AORED protocol offers significant improvement in network stability and network life time as compared to DEEC and SEP.
As future work, the same idea of using CDS in AORED can be evaluated by making a certain number of WSN nodes mobile. Furthermore, various other algorithms of CDS construction can be explored by coupling them with the same AORED protocol.

acknowledgment The authors would like to thank Ashab Tariq for his help in formatting the manuscript.

References

[1] M. J. Handy, M. Haase, and D. Timmermann, "Low energy adaptive clustering hierarchy with deterministic cluster-head selection," in *Mobile and Wireless Communications Network, 2002. 4th International Workshop on*, pp. 368–372, 2002.

[2] G. Smaragdakis, I. Matta, and A. Bestavros, "SEP: A Stable Election Protocol for clustered heterogeneous wireless sensor networks," in *Second International Workshop on Sensor and Actor Network Protocols and Applications (SANPA 2004)*, (Boston, MA), August 2004.

[3] L. Qing, Q. Zhu, and M. Wang, "Design of a distributed energy-efficient clustering algorithm for heterogeneous wireless sensor networks," *Comput. Commun.*, vol. 29, pp. 2230–2237, Aug. 2006.

[4] H. Lee, M. Jang, and J.-W. Chang, "A new energy-efficient cluster-based routing protocol using a representative path in wireless sensor networks," *International Journal of Distributed Sensor Networks*, vol. 10, no. 7, 2014.

[5] D. b. Zou and Y. B. Wang, "Adaptive energy-aware routing framework in transmission cost constrained wireless sensor networks," in *2013 IEEE Global Communications Conference (GLOBECOM)*, pp. 534–538, Dec 2013.

[6] C. Hsu, H. Chu, and J. Liaw, "Connectivity and energy-aware clustering approach for wireless sensor networks," in *2014 IEEE International Conference on Systems, Man, and Cybernetics, SMC 2014, San Diego, CA, USA, October 5-8, 2014*, pp. 1708–1713, 2014.

[7] M. Natarajan, R. Arthi, and K. Murugan, "Energy aware optimal cluster head selection in wireless sensor networks," in *Computing, Communications and Networking Technologies (ICCCNT),2013 Fourth International Conference on*, pp. 1–4, July 2013.

[8] M. Tripathi, R. B. Battula, M. S. Gaur, and V. Laxmi, "Energy efficient clustered routing for wireless sensor network," in *Mobile Ad-hoc and Sensor Networks (MSN), 2013 IEEE Ninth International Conference on*, pp. 330–335, IEEE, 2013.

[9] T. H. Chiang and J. S. Leu, "Regional energy aware clustering with isolated nodes in wireless sensor networks," in *2014 IEEE 25th Annual International Symposium on Personal, Indoor, and Mobile Radio Communication (PIMRC)*, pp. 1829–1833, Sept 2014.

[10] R. D. Gawade and S. Nalbalwar, "A centralized energy efficient distance based routing protocol for wireless sensor networks," *Journal of Sensors*, vol. 2016, 2016.

[11] J. Singh, B. P. Singh, and S. Shaw, "A new leach-based routing protocol for energy optimization in wireless sensor network," in *Computer and Communication Technology (ICCCT), 2014 International Conference on*, pp. 181–186, Sept 2014.

[12] A. E. Tand M. G, "An improved leach protocol for indoor wireless sensor networks," in *Signal Processing and Integrated Networks (SPIN), 2014 International Conference on*, pp. 432–437, Feb 2014.

[13] A. Antoo and A. R. Mohammed, "Eem-leach: Energy efficient multi-hop leach routing protocol for clustered wsns," in *Control, Instrumentation, Communication and Computational Technologies (ICCICCT), 2014 International Conference on*, pp. 812–818, July 2014.

[14] R. K. Kodali, V. S. K. A., S. Bhandari, and L. Boppana, "Energy efficient m- level leach protocol," in *Advances in Computing, Communications and Informatics (ICACCI), 2015 International Conference on*, pp. 973–979, Aug 2015.

[15] A. Bharti, C. Devi, and V. Bhatia, "Enhanced energy efficient leach (eee-leach) algorithm using mimo for wireless sensor network," in *2015 IEEE International Conference on Computational Intelligence and Computing Research (ICCIC)*, pp. 1–4, Dec 2015.

[16] A. Razaque, S. Mudigulam, K. Gavini, F. Amsaad, M. Abdulgader, and G. S. Krishna, "H-leach: Hybrid-low energy adaptive clustering hierarchy for wireless sensor networks," in *2016 IEEE Long Island Systems, Applications and Technology Conference (LISAT)*, pp. 1–4, April 2016.

[17] P. K. Batra and K. Kant, "Leach-mac: A new cluster head selection algorithm for wireless sensor networks," *Wirel. Netw.*, vol. 22, pp. 49–60, Jan. 2016.

[18] E. Shakshuki, W. Akkari, B. Bouhdid, and A. Belghith, "The 6th international conference on ambient systems, networks and technologies (ant-2015), the 5th international conference on sustainable energy information technology (seit-2015) leatch: Low energy adaptive tier clustering hierarchy," *Procedia Computer Science*, vol. 52, pp. 365 – 372, 2015.

[19] A. Razaque, M. Abdulgader, C. Joshi, F. Amsaad, and M. Chauhan, "P-leach: Energy efficient routing protocol for wireless sensor networks," in *2016 IEEE Long Island Systems, Applications and Technology Conference (LISAT)*, pp. 1–5, April 2016.

[20] S. Lindsey and C. S. Raghavendra, "Pegasis: Power-efficient gathering in sensor information systems," in *Aerospace Conference Proceedings, 2002. IEEE*, vol. 3, pp. 3–1125–3–1130 vol.3, 2002.

[21] I. Sohn, J. H. Lee, and S. H. Lee, "Low-energy adaptive clustering hierarchy using affinity propagation for wireless sensor networks," *IEEE Communications Letters*, vol. 20, pp. 558–561, March 2016.

[22] S. Yadav and R. S. Yadav, "A review on energy efficient protocols in wireless sensor networks," *Wirel. Netw.*, vol. 22, pp. 335–350, Jan. 2016.

[23] P. K. Batra and K. Kant, "Stable cluster head selection in leach protocol: A cross-layer approach," in *Proceedings of the 7th ACM India Computing Conference*, COMPUTE '14, (New York, NY, USA), pp. 15:1–15:6, ACM, 2014.

[24] N. Zaman, L. T. Jung, and M. M. Yasin, "Enhancing energy efficiency of wireless sensor network through the design of energy efficient routing protocol," *Journal of Sensors*, vol. 2016, 2016.

[25] Y. Wu, *Connected Dominating Set Construction and Application in Wireless Sensor Networks*. PhD thesis, Atlanta, GA, USA, 2010. AAI3405735.

[26] A. H. Karbasi and R. E. Atani, "Application of dominating sets in wireless sensor networks," *Int. J. Sec. Appl*, vol. 7, no. 4, 2013.

[27] J. P. Mohanty, C. Mandal, C. Reade, and A. Das, "Construction of minimum connected dominating set in wireless sensor networks using pseudo dominating set," *Ad Hoc Netw.*, vol. 42, pp. 61–73, May 2016.

[28] D. Djenouri and M. Bagaa, "A variant of connected dominating set in unit disk graphs for applications in communication networks," in *2015 IEEE International Conference on Electro/Information Technology (EIT)*, pp. 457–461, May 2015.

[29] J. Zhang, L. Xu, S. M. Zhou, and W. Wu, "Constructing connected dominating set based on crossed cube in wsn," in *Intelligent Networking and Collaborative Systems (INCoS), 2013 5th International Conference on*, pp. 443–447, Sept 2013.

[30] P. S. Vinayagam, "A survey of connected dominating set algorithms for virtual backbone construction in ad hoc networks," *International Journal of Computer Applications*, vol. 143, pp. 30–36, Jun 2016.

[31] F. Dai and J. Wu, "An extended localized algorithm for connected dominating set formation in ad hoc wireless networks.," *IEEE Trans. Parallel Distrib. Syst.*, vol. 15, no. 10, pp. 908–920, 2004.

Adaptive Error Detection Method for P300-based Spelling Using Riemannian Geometry

Attaullah Sahito, M. Abdul Rahman, Jamil Ahmed
Department of Computer Science
Sukkur Institute of Business Administration
Airport road Sukkur, Pakistan

***Abstract*—Brain-Computer Interface (BCI) systems have become one of the valuable research area of ML (Machine Learning) and AI based techniques have brought significant change in traditional diagnostic systems of medical diagnosis. Specially; Electroencephalogram (EEG), which is measured electrical activity of the brain and ionic current in neurons is result of these activities. A brain-computer interface (BCI) system uses these EEG signals to facilitate humans in different ways. P300 signal is one of the most important and vastly studied EEG phenomenon that has been studied in Brain Computer Interface domain. For instance, P300 signal can be used in BCI to translate the subject's intention from mere thoughts using brain waves into actual commands, which can eventually be used to control different electro mechanical devices and artificial human body parts. Since low Signal-to-Noise-Ratio (SNR) in P300 is one of the major challenge because concurrently ongoing heterogeneous activities and artifacts of brain creates lots of challenges for doctors to understand the human intentions. In order to address above stated challenge this research proposes a system so called Adaptive Error Detection method for P300-Based Spelling using Riemannian Geometry, the system comprises of three main steps, in first step raw signal is cleaned by preprocessing. In second step most relevant features are extracted using xDAWN spatial filtering along with covariance matrices for handling high dimensional data and in final step elastic net classification algorithm is applied after converting from Riemannian manifold to Euclidean space using tangent space mapping. Results obtained by proposed method are comparable to state-of-the-art methods, as they decrease time drastically; as results suggest six times decrease in time and perform better during the inter-session and inter-subject variability.**

***Keywords*—Brain Computer Interface; EEG; P300; Riemannian geometry; xDAWN; Covariances; Tangent Space; Elastic net**

I. Introduction

In recent ML (Machine Learning) techniques have been vastly used to solve the various medical classification problems such as Liver, Heart, Neurological disorders and others. Particularly; human intuition interpretation using P-300 signals from EEG (Electroencephalogram) have become one of the significant problems, since BCI (Brain Computer Interface) establishes point of communication between doctors and patients where the subject is found with complete paralyzed situation. For example, locked-in syndrome paralysis is a condition in which patient is found awake and fully aware but no expression could be made by patients to communicate with doctors. The patients with spinal cord injury leading to ALS (Amyotrophic Lateral Sclerosis) so called neuro-degenerative disease in which loss of voluntary control muscles produces human paralysis and patient become unable to communicate. In these conditions Brain Computer Interface can help in controlling of external devices or as a reliable tool for communication. Electrical activity of neurons in brain can be recorded using Electroencephalogram (EEG). This EEG data is of great use: neurologists can use for diagnosing neurological disorders. BCI experts can use for means of communication. Nowadays BCI has become one of assistive tool for human beings, it can control wheel chairs, video games, entertainment, spellers, and other assistive technologies [1]. In order to address above stated challenge this paper proposes a system called Adaptive Error Detection method for P300-Based Spelling using Riemannian Geometry, the system comprises of three main steps, in first step raw signal is cleaned by preprocessing. In second step most relevant features are extracted using xDAWN spatial filtering along with covariance matrices for handling high dimensional data and in final step elastic net classification algorithm is applied after converting from Riemannian manifold to Euclidean space using tangent space mapping. Since BCI is vastly used for controlling external devices and as a reliable tool for communication. In these circumstances as stated above; BCI can be used to establish a reliable communication channel directly from patients brain signals to the computer [2], but low Signal-to-Noise Ratio (SNR) in P300 is one of the major challenges because concurrently ongoing heterogeneous activities and artifacts of brain create lots of difficulties for doctors to understand the human intentions.

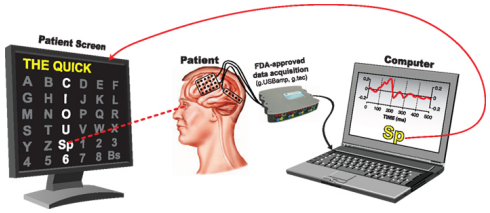


Fig. 1. P300 Speller

P300 signal is one of the vastly explored non-invasive Brain Computer Interface scenario. Response of P300 signal can be considered as an oddball scenario, where low probability of desired events is mixed with high probability of undesired events. An ERP or event-related-potential can be described as recorded P300 response during the course of decision making when subject reacts to different kinds of stimuli. P300 [3] speller has 6X6 matrix of characters containing alpha-numeric characters as shown here in Figure. 1.

Different kinds of stimuli can be explored such as auditory,

visual etc., however P300 speller uses visual stimuli [4] only. In P300 Speller scenario a cap containing electrodes, is placed on scalp of subject. These electrodes record brain waves data produced against a visual stimuli of letter displayed on screen. Raw EEG signal is inherently very noisy; eventually, it has very low SNR. Therefore, it is very hard to differentiate true P300 response from recorded EEG signal. Low SNR is due to different ongoing electrical activities of neurons in brain while performing other physiological tasks simultaneously. This noise poses a challenge of extracting true ERP and interpreting brain activity correctly. Problem is to detect errors during spelling task made by P300 speller, given subject's brain waves. Machine Learning (ML) algorithms play crucial role in formulating Brain Computer Interface systems. For example, ML algorithms can help in extracting influential and most relevant features from data, because recorded signals are inherently noisy due to many reasons. Feature extraction can be helpful in removing artifacts, noise and other undesired factors from recorded data. Classification algorithm is used to set learning parameters on training data and employed on test data. In this scenario, goal of classifier is to predict based on training data; whether P300 speller has recognized Target letter correctly or not according to subject's intention. In this paper, we are proposing a ML pipeline for classification of EEG data from P300 speller having robust and reliable feature extraction method and classification algorithm, which minimizes misinterpreted classified commands against user's actual desired intentions and provide reliable means of communication. For this objective, BCI system must possess two key properties:

1) Better classification technique for P300 Speller which enhances across-subject generalization and across-session generalization, having optimal accuracy.
2) Fast convergence to optimal parameters requiring minimal calibration and training data.

This paper is organized in five sections. Section one is used to describe the introduction of this paper, section two describes the related works, section three defines the materials and methods, section four represents the results and in section five conclusion and discussion is discussed.

A. Related Work

Classification of P300 speller is of one the active research area(s) of ML (Machine Learning) and many approaches have been proposed to solve the classification problem. Our approach deals as productive modelling for P300 classification in machine learning domain. Some of the related works have been presented as below.

For P300 classification, comparison of different classifier algorithms is presented in [6], between different classifiers such as stepwise linear discriminant analysis (SWLDA), support vector machines (SVMs) and LDA (Linear Discriminant Analysis). Results suggest that SWLDA performs better than other classifiers. Although SWLDA works as simple LDA, but at start SWLDA has no feature in discrimination function; it starts adding features one by one using their statistical significance from p-value. In this paper only 16 electrodes were used. This paper proposes a scalable system, which can incorporate as many as 56 electrodes.

A system [7] was proposed to classify ERP data from P300 experiment with optimal accuracy. Spatial as well as spatio-temporal features are used along with Linear Discriminant Analysis with shrinkage as classifier and proposed an analytical method for estimating regularization parameter for LDA. Then comparison between different classifiers such as SWLDA [6] and simple LDA is presented. Results suggest their model performs better than other two. They used 55 channels but for only 7-time sample points after occurrence of an event. Epoch window size of 7-time sample points is very small; it is likely that important temporal information for whole trial can be missed. This paper proposes a scalable system, which can incorporate as many as 260 time points, maintaining full spatial as well as temporal information.

Comparison of different feature selection and classification methods is presented by [8] using five different datasets from Brain Computer Interface paradigm. For model building; spectral, spatial and spatio-temporal filtering as feature selection was applied and for classification four methods LDA, SWLDA, rLDA (regularized LDA) and rLR (regularized logistic regression) were used. Results suggest that regularized classifiers are better in terms of performance. This suggested approach requires a lot of manual parameter tuning and results are not generalizable across sessions and across users.

System based on Riemannian geometry was firstly introduced by Congedo et al. [9] for classification of BCI systems. Data from all three modalities of BCI Motion Imagery (MI), Steady State Visually Evoked Potential (SSVEP) and ERP from 22 electrodes. They proposed a classifier known as MDM (Minimum Distance to Mean) classifier, which works on principle of Riemannian geometry. This classifier works by calculating covariance mean matrices for each class in training data as a representative and assigns the label to test data by calculating their distances from mean covariance matrices of classes. For comparison they used Common Spatial Patterns+ [10] + Linear Discriminant Analysis [11]. Results suggest that, Riemannian geometry based classifier Minimum Distance to Mean (MDM) works better and requires small amount of data for training and also works better in across-subject and across-session generalization. As robust and efficient classification algorithms like Neural network, SVM, etc. work in euclidean space; so cannot be applied to covariance matrices belonging to riemannian manifold directly, this is main limitation of this approach.

Feature selection technique known as xDAWN based on spatial filtering was introduced by [12]. Main motivation behind this technique was to discriminate and enhance P300 evoked response. This technique works on assumption that P300 response related to an ERP is least frequent and target response occurs simultaneously, has very little spatial subspace span in the recorded signal. This algorithm is based on QR factorization of matrices [13]. For comparison, 3 subjects data having 29 electrodes were used. Results comparison was done between different feature extraction techniques such as xDAWN spatial filtering, ICA and PCA, and classification using Bayesian Linear Discriminant Analysis (BLDA). Results suggest that xDAWN spatial filtering technique has brought significant improvement on accuracy as compared with ICA and PCA. However, if xDAWN components are increased, then performance slightly decreases and computational complexity

increases, this is main limitation of algorithm.

As discussed earlier famous classification techniques works in euclidean space and hence cannot be applied directly in Riemannian Manifolds. To overcome this limitation [14],[15] introduced Tangent Space mapping (TS). Tangent Space bridges euclidean space and Riemannian manifolds. Tangent space mapping projects covariance matrices belonging to Riemann manifold into Euclidean space vectors. By using this mapping, one can use classical and efficient classifiers such as LDA, SVM etc. on covariance matrices directly, instead of using MDM [9]. Results suggest that, significant improvement can be achieved; without need of spatial filtering of electrodes as they are exploited in their native space. Although, from BCI; Motor Imagery (MI) was considered in these both papers [14],[15], though this technique can be extended for classification of P300 signal.

All above stated approaches have performed better, but some techniques work well on one data set while perform poorly on other data set. However, [5],[9], [16] proposed a system that uses Riemannian based covariance matrices as features along with other additional meta features for instance word length, feedback detail, etc. to enhance the prediction performance. These newly introduced techniques has brought significant improvement for classification, but also brought adverse effect on computational complexity of the classifier. Hence, P300 classification system direly needs a novel approach which can minimize computational complexity along with consideration of optimal features towards time accuracy trade-off.

II. Material and Methods

Proposed approach is based on modelling machine learning pipeline, which deals with the classification problem of the P300 speller; whether an event is erroneous or not. Our proposed methodology as graphically shown in Figure. 2, is divided into four interconnected steps where each step is assigned several inputs to interact with each other. In first step; raw data is cleaned using simple preprocessing techniques, in second step BCI related features are extracted from cleaned data using xDAWN spatial filters along with covariance matrices and tangent space mapping. In third step classification model is constructed by using elastic net and in fourth step the classification accuracy have been measured. These steps are discussed in detail below.

A. Preprocessing

In preprocessing step, EOG(ElectroOculoGram) channel is removed which gives information about the noise caused by blinking of subjects eye. 5^{th} order butter-worth filter is employed to bandpass rest EEG channels between 1 to 40 Hz. Butter-worth filter is maximally flat magnitude filter, which have response in its passband. For each trial each epoch window is set to 1.3 seconds after the occurrence of feedback event. After preprocessing of raw signals, features are extracted using below defined ML pipeline.

B. Feature Extraction

Figure. 2 gives an overview of all preprocessing steps, feature extraction techniques and model building steps. Our

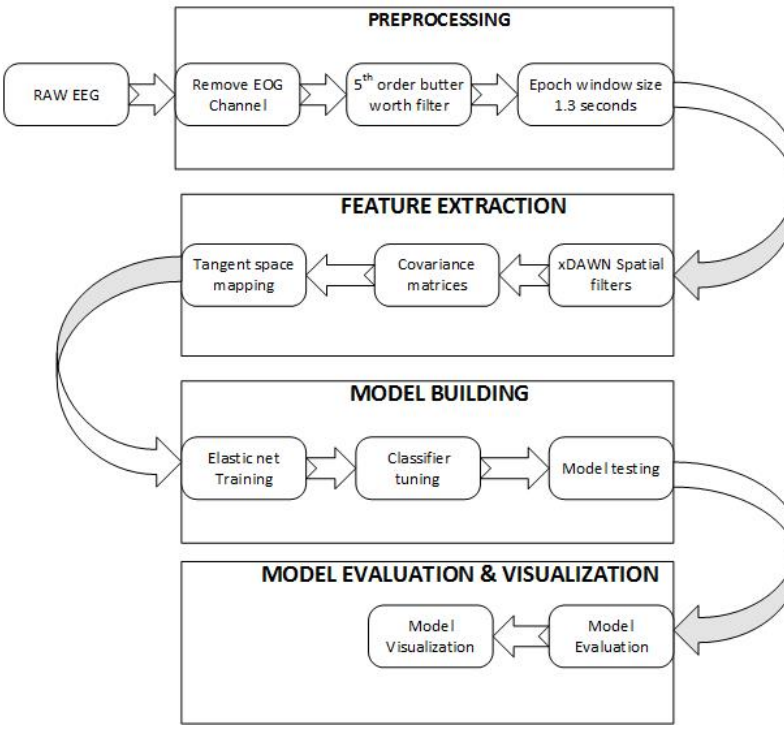


Fig. 2. Adaptive error detection system for P300 speller workflow

feature extraction step comprises of three interconnected sub steps Xdawn spatial filters, covariances matrices and tangent space mapping. All these are discussed in detail as under.

1) xDAWN: As discussed earlier, brain waves recorded through EEG are noisy, this recorded EEG signal has information about true P300 signal as well as other ongoing background activities of the brain such as muscular artifacts, eye blinking etc. Therefore, acquired EEG data has very mcuh noise and this low SNR makes classification task challenging. Different techniques have been investigated in literature to improve EEG signal, spatial filtering one of them. For instance, independent component analysis (ICA) was used [17] to enhance SNR. Limitation of such techniques are that, they are not designed specifically for Brain Computer Interface. To overcome this limitation [12] an unsupervised spatial filtering technique known as xDAWN was introduced. This technique works on assumption: P300 related signal is less frequent and has smaller synchronous subspace in the recorded EEG signal. These synchronous responses for each channel are calculated and then spatial filters are calculated using these responses, in order to make true P300 signals standout form other artifacts and the noise.

Let us $X \in R^{N_t x N_s}$ denotes actual recorded raw EEG signals. In which each entry $(i,j)^{th}$ is $x_j(i)$, which represents data from j^{th} channel at instant of time i. Here N_s represents index of channel and N_t represents time index of a trail. Let us assume $a_j(t)$ denote the Event Related Potential signal from j^{th} electrode channel at time course of index t for that trail, whereas $A \in R^{N_e x N_s}$ denotes Event Related Potential signal, where $(i,j)^{th}$ value corresponds to $a_j(i)$. Here N_s represents index of channel and N_e is index of time point for Event Related Potential which here corresponds to 260 which is given

by single trial duration of 1.3 seconds. Recorded signal of EEG can be formulated as combination of Event Related Potential A and Noise N:

$$X = DA + N \tag{1}$$

Where D represents Toeplitz matrix $D \in R^{N_e x N_s}$. Solution of (1) is given by QR factorization [13].

2) Covariances: As discussed above Signal to noise Ratio (SNR) can be enhanced by spatial filtering for better design of Brain Computer Interface systems, but it requires substantial amount of data for training purpose, therefore for BCI scenario more repeated trials from the subject are needed, this makes spatial filtering soleley unfeasible choice for on-line scenario. For BCI systems, covariance matrices for feature extraction were introduced by [15], further [16] improved covariance matrices by including temporal information as well. This all has been possible due to recent theoretical advancements in field of Information geometry [18]. The information geometry is a field of mathematics, which takes probability distributions as points of a Riemannian manifold (Manifold has resemblance to homemorphic euclidean space near each point.). Nowadays it is widely applied in various fields. For example, processing of radar signals [19], diffusion tensor [20] and digital image processing [21]. Processing the covariance matrices in their native manifold has added benefit. Sample covariance matrix (SCM) also known as prototyped ERP response matrix is calculated for each class by taking average of all epochs belonging to same class. Let $P_1 \in R^{s*t}$ be the sample covariance matrix (SCM) for class 1.

$$P_1 = \frac{1}{I}\sum_{i \in I} X_i \tag{2}$$

Where I is the indexes of i^{th} trail. For each trial x_i, modified trail $\tilde{x}_i$ is given by combining it with SCM:

$$\tilde{x_i} = \begin{bmatrix} P1 \\ X_i \end{bmatrix}$$

Final covariance matrices are computed using SCM:

$$\tilde{\sum} i = \frac{1}{N-1}\tilde{X}_i \tilde{X}_i{}^T \tag{3}$$

The resultant covariance matrix of each epoch is concatenated with xDAWN spatial filters and passed to next step. One can directly only apply Minimum Distance to Mean (MDM) classifier on covariance matrices [9], [16] because they belong to Riemannian manifold of symmetric positive definite matrices (SPD) [22].

3) Tangent Space: As discussed earlier [9,16] that we cannot apply classical classification algorithms to covariance matrices without modification, because they are from riemannian manifold. Although [15] used support vector machine (SVM) by modifying kernel, but this is not an obvious choice. Tangent space mapping helps us in using robust and efficient classifiers for instance svm, elastic net etc. easily by converting riemannian covariance matrices to euclidean space. Tangent vectors for euclidean space are calculated using tangent space mapping on covariance matrices belonging to riemannian manifold. Process of tangent space mapping is geometrically shown in Figure. 3.

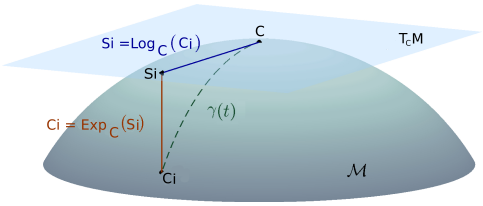


Fig. 3. Tangent space TcM corresponds to manifold M at point C. Si tangent vectors are estimated using Logarithmic map LogC(:). The Exponential-map ExpC(:) is used for invers mapping [15].

Logarithmic mapping is used to calculate tangent vectors Si of tangent space TcM, for each point C (which here is actually covariance matrices) of manifold M as shown in Figure. 3. Tangent vector space TcM for that point in euclidean space, which is locally homomorphic to manifold can be calculated using:

$$S_i = Log_C(C_i) = C^{\frac{1}{2}} log(C^{\frac{-1}{2}} C_i C^{\frac{-1}{2}}) C^{\frac{1}{2}} \tag{4}$$

C. Model building

Different classifier models were used, but elastic net performed better than others. Elastic net learning algorithm is widely used to predict in different classification problems. It is one of the linear regularized regression algorithm, which addresses shortcomings of lasso as well as ridge regression by introducing l1 and l2 penalty [23] and also handles numerical data very well. After feature extraction, elastic net is trained on training set comprises of 5440 trails, after tangent space mapping while p = 2211 number of predictors, class labels of training data is represented by $Y = (y_1, y_2, y_3, ...y_n)^T$ and $X = [X_1]...[X_p]$ be model matrix where $x_j = (x_{1j}, ...x_{nj})^T$. Let λ_1 and λ_2 be the fixed non-negative integers, then objective function for elastic net algorithm in terms of minimization can be defined as:

$$L(\lambda_1, \lambda_2, \beta) = |Y - X\beta|^2 + \lambda_2|\beta|^2 + \lambda_1|\beta|_1 \tag{5}$$

For training and testing purpose Elastic net is used. For tuning of hyper parameters of elastic net were tuned by cross validation. Prediction is done using conventional linear regression equation as:

$$\hat{y} = \beta_0 + x_1\beta_1 + ... + x_p\beta_p \tag{6}$$

D. Model evaluation

In performance evaluation, area under the ROC (Receiver Operating Characteristic) curve (AUC) is used, as it has already been used widely for binary classification tasks [24]. AUC is calculated using:

$$AUC = \frac{1}{N_+ + N_-}\sum_{i=1}^{N_+}\sum_{j=1}^{N_-} I(x_i > y_j) \tag{7}$$

Where N+ denotes frequency of instances belonging to positive class and N- denotes frequency of instances belonging to negative class. While $x_1, \ldots, x_{N+}$ are probability scores predicted by model for N_+ positives and $y_1, \ldots, y_{N-}$ denotes probability scores predicted by model for N_- negative class.

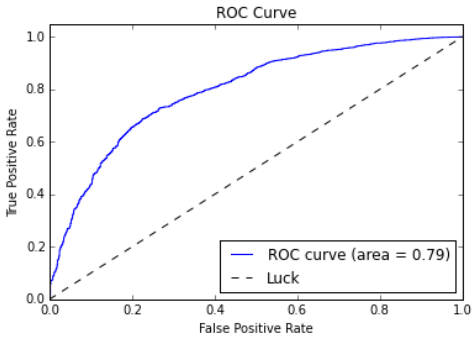


Fig. 4. ROC Curve estimated for P300 classification using proposed system

On other hand values of x_i and y_j are normalized between $(0, 1)$.

For model building and evaluation Data set is used from an experiment [25]. In this experiment, brain activity was recorded with total 56 passive EEG electrode channels. Data was collected from 26 subjects (16 training subjects and 10 test subjects) during 5 different spelling sessions. All subjects participated in five sessions, only 12 five-letter words were spelt in first four sessions by subject and in last session 20 five-letter words were spelt. For each letter, trail is lasted for 1.3 seconds. All subjects in these five sessions constitute total training 5440 trials and 3400 test trials.

III. Results

In figure. 4 presents the results of fully automated process after applying proposed ML pipeline using ROC curve having 0.79 AUC (Area under curve). While performance for different users across different sessions is shown in [Table 1] in terms of AUC. Figure. 5 compares the AUC obtained by proposed method with state-of-the-art and award winning method by Alexandre [5], which shows almost equivalent AUC obtained by both methods, this is also supported by figure. 7; where samples of size 200, 500, 800 and 1200 were used from test data and AUC is compared for both models. Another comparison is performed by considering time as a variable. Figure. 6 shows comparison between time taken by proposed approach and Alexandres [5] approach, which shows huge difference in time taken by proposed approach and Alexandres approach; on average proposed approach takes 37 seconds while Alexandres [5] approach took 224 seconds. This result is also supported by results presented in figure. 8; where time taken by both approaches for samples of size 200, 500, 800 and 1200 from test data is compared.

IV. Conclusion

Due to low Signal to Noise Ratio and high dimensional nature of EEG data, classification of P300 signal is one of the challenging problem. Although various approaches have been reported in the literature for classification of P300 signals based on feature extraction methods but most of them uses spatial filters [11] and exploits riemannian geometry [9] but classification of P300 data requires significant method to preprocess. This paper proposes a machine learning pipeline from preprocessing to construction of classification model for

TABLE I. Overall Performance of Proposed Methodology

Session→	1	2	3	4	5
Subject 1	0.79	0.80	0.81	0.66	0.56
Subject 3	0.84	0.88	0.69	0.75	0.67
Subject 4	0.80	0.94	0.91	0.90	0.80
Subject 5	0.61	0.70	0.65	0.52	0.65
Subject 8	0.91	0.90	0.90	0.86	0.75
Subject 9	0.63	0.75	0.71	0.87	0.75
Subject 10	0.90	0.96	0.90	0.93	0.96
Subject 15	0.94	0.79	0.76	0.99	0.82
Subject 19	0.64	0.58	0.63	0.65	0.69
Subject 25	0.83	0.81	0.64	0.72	0.65

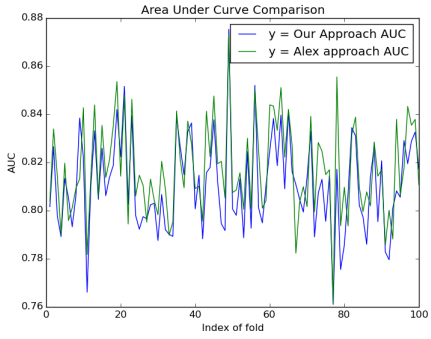


Fig. 5. Comparing Area Under Curve with Alexandre Model

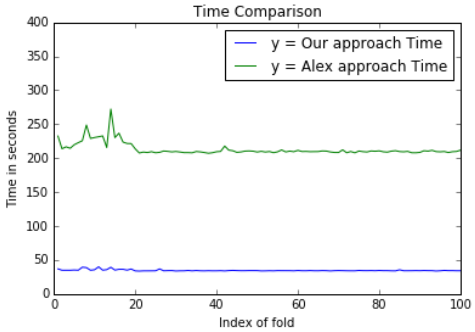


Fig. 6. Comparing time taken (seconds) with Alexandre Model

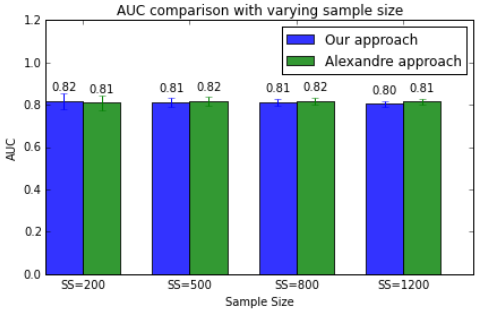


Fig. 7. AUC comparison with varying sample size

P300 data, by using tangent space mapping from riemannian geometry while same kind of approaches were also applied in Motion Imagery (another modality of Brain Computer Interface). System proposed in this study is comprised of three steps. In first step, raw EEG signals form brain are

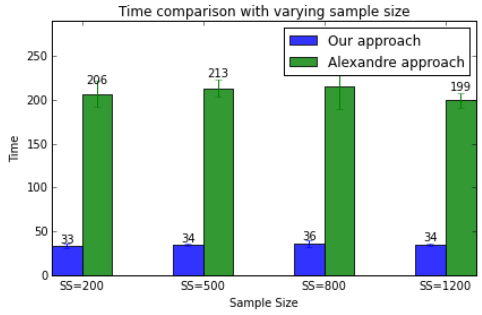


Fig. 8. Time comparison with varying sample size

preprocessed by removing irrelevant electrodes and filtering through 5th order butter-worth filter and discarding signals are bandpass 1-40 Hz. In second step, feature extraction is performed by estimating five xDAWN spatial filters and concatenating with covariances matrices and then tangent space mapping is applied to intermediate results for projecting xDAWN covariances from riemannian manifold into homologous euclidean space. In final step, classification using Elastic net is performed; whose parameters are tuned by cross validation. Based on presented results, it can be concluded that the system classifies the P300 signals efficiently and achieves better AUC while low computational complexity for most of the cases. Also there is drastic decrease in time taken to build a model while maintaining higher AUC while supplying low amount of training data. Proposed model shows better performance during inter-session and inter-subject variability and accomplishes comparable AUC 0.79 in just 37 seconds as compared against state-of-the-art approach, having AUC 0.80 in 224 seconds.

This study can be extended in different ways. As this paper only linear classifiers were used, but other complex classifiers such as neural networks can be used to improve accuracy. Furthermore, we used all electrodes for classification. However, one would like to minimize and select only most relevant electrodes for classification.

REFERENCES

[1] J d R Millán, Rüdiger Rupp, Gernot R Müller-Putz, Roderick Murray-Smith, Claudio Giugliemma, Michael Tangermann, Carmen Vidaurre, Febo Cincotti, Andrea Kübler, Robert Leeb, et al. Combining brain–computer interfaces and assistive technologies: state-of-the-art and challenges. Frontiers in neuroscience, 4, 2010.

[2] Thorsten O Zander, Christian Kothe, Sebastian Welke, and Matthias Rötting. Utilizing secondary input from passive brain-computer interfaces for enhancing human-machine interaction. In Foundations of Augmented Cognition. Neuroergonomics and Operational Neuroscience, pages 759–771. Springer, 2009.

[3] Lawrence Ashley Farwell and Emanuel Donchin. Talking off the top of your head: toward a mental prosthesis utilizing event-related brain potentials. Electroencephalography and clinical Neurophysiology, 70(6): 510–523, 1988.

[4] A Furdea, S Halder, DJ Krusienski, D Bross, F Nijboer, N Birbaumer, and A Kübler. An auditory oddball (p300) spelling system for brain-computer interfaces. Psychophysiology, 46(3):617–625, 2009.

[5] Alexandre Barachant, Rafał Cycon, and Cedric Gouy-Pailler. P300-speller: Géométrie riemannienne pour la détection multi-sujets de potentiels d'erreur. In GRETSI 2015, 2015.

[6] Dean J Krusienski, Eric W Sellers, François Cabestaing, Sabri Bayoudh, Dennis J McFarland, Theresa M Vaughan, and Jonathan R Wolpaw. A comparison of classification techniques for the p300 speller. Journal of neural engineering, 3(4):299, 2006.

[7] Benjamin Blankertz, Steven Lemm, Matthias Treder, Stefan Haufe, and Klaus-Robert Müller. Single-trial analysis and classification of erp componentsa tutorial. NeuroImage, 56(2):814–825, 2011.

[8] Jason Farquhar and N Jeremy Hill. Interactions between pre-processing and classification methods for event-related-potential classification. Neuroinformatics, 11(2):175–192, 2013.

[9] Marco Congedo, Alexandre Barachant, and Anton Andreev. A new generation of brain-computer interface based on riemannian geometry. arXiv preprint arXiv:1310.8115, 2013.

[10] Moritz Grosse-Wentrup and Martin Buss. Multiclass common spatial patterns and information theoretic feature extraction. Biomedical Engineering, IEEE Transactions on, 55(8):1991–2000, 2008.

[11] Fabien Lotte and Cuntai Guan. Regularizing common spatial patterns to improve bci designs: unified theory and new algorithms. Biomedical Engineering, IEEE Transactions on, 58(2):355–362, 2011.

[12] Bertrand Rivet, Antoine Souloumiac, Virginie Attina, and Guillaume Gibert. xdawn algorithm to enhance evoked potentials: application to brain–computer interface. Biomedical Engineering, IEEE Transactions on, 56(8):2035–2043, 2009.

[13] Gene H Golub and Charles F Van Loan. Matrix computations, volume 3. JHU Press, 2012.

[14] Alexandre Barachant, Stéphane Bonnet, Marco Congedo, and Christian Jutten. Multiclass brain–computer interface classification by riemannian geometry. Biomedical Engineering, IEEE Transactions on, 59(4):920–928, 2012.

[15] Alexandre Barachant, Stéphane Bonnet, Marco Congedo, and Christian Jutten. Classification of covariance matrices using a riemannian-based kernel for BCI applications. Neurocomputing, 112:172–178, 2013.

[16] Alexandre Barachant and Marco Congedo. A plug&play p300 bci using information geometry. arXiv preprint arXiv:1409.0107, 2014.

[17] Neng Xu, Xiaorong Gao, Bo Hong, Xiaobo Miao, Shangkai Gao, and Fusheng Yang. Bci competition 2003-data set iib: enhancing p 300 wave detection using ica-based subspace projections for bci applications. IEEE transactions on biomedical engineering, 51(6):1067–1072, 2004.

[18] Shun-ichi Amari and Hiroshi Nagaoka. Methods of information geometry, volume 191 of translations of mathematical monographs. American Mathematical Society, page 13, 2000.

[19] F Barbaresco. Innovative tools for radar signal processing based on cartans geometry of spd matrices & information geometry. In Radar Conference, 2008. RADAR'08. IEEE, pages 1–6. IEEE, 2008.

[20] P Thomas Fletcher and Sarang Joshi. Principal geodesic analysis on symmetric spaces: Statistics of diffusion tensors. In Computer Vision and Mathematical Methods in Medical and Biomedical Image Analysis, pages 87–98. Springer, 2004.

[21] Oncel Tuzel, Fatih Porikli, and Peter Meer. Pedestrian detection via classification on riemannian manifolds. Pattern Analysis and Machine Intelligence, IEEE Transactions on, 30(10):1713–1727, 2008.

[22] Maher Moakher. A differential geometric approach to the geometric mean of symmetric positive-definite matrices. SIAM Journal on Matrix Analysis and Applications, 26(3):735–747, 2005.

[23] Hui Zou and Trevor Hastie. Regularization and variable selection via the elastic net. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 67(2):301–320, 2005.

[24] Shaomin Wu and Peter Flach. A scored auc metric for classifier evaluation and selection. In Second Workshop on ROC Analysis in ML, Bonn, Germany, 2005.

[25] Perrin Margaux, Maby Emmanuel, Daligault Sébastien, Bertrand Olivier, and Mattout Jérémie. Objective and subjective evaluation of online error correction during p300-based spelling. Advances in Human-Computer Interaction, 2012:4, 2012.

Evaluation of OLSR Protocol Implementations using Analytical Hierarchical Process (AHP)

Ashfaq Ahmad Malik and Tariq Mairaj Rasool Khan
PN Engineering College (PNEC)
National University of Science and Technology (NUST)
Karachi, Pakistan

Athar Mahboob
Khwaja Fareed University of
Engineering and Information Technology (KFUEIT)
Rahim Yar Khan, Pakistan

***Abstract*—Adhoc networks are part of IEEE 802.11 Wireless LAN Standard also called Independent Basic Service Set (IBSS) and work as Peer to Peer network by default. These work without the requirement of an Infrastructure (such as an Access Point) and demands specific routing requirements to work as a multi-hop network. There are various Adhoc network routing protocols which are categorized as Proactive, Reactive and Hybrid. OLSR (a proactive routing protocol) is one of widely used routing protocols in adhoc networks. In this paper an empirical study and analysis of the various OLSR implementations (by different research groups and individuals) has been conducted in light of Relative Opinion Scores (ROS) and Analytical Hierarchical Process (AHP) Online System software. Based on quantitative comparison of results, it is concluded that OLSRd project is most updated and best amongst six variants of OLSR protocol implementations.**

***Keywords*—OLSR; MANET; AHP; Routing Protocols**

I. Introduction and Background Information

We are working on Mobile Adhoc Networks (MANETs) based on 802.11 WLAN Standard based mobile devices to build a trustworthy collaborative system. Due to peculiar nature of adhoc networks (as compared to infrastructure networks); the normal routing protocols used in infrastructure or Access Point (AP) based network can not be directly applied or used in adhoc networks. There are many routing protocols defined and used as an outcome of research in adhoc networks such as Adhoc On-demand Distance Vector (AODV), Fisheye State Routing (FSR) Protocol, Destination-Sequenced Distance-Vector (DSDV), Optimized Link State Routing (OLSR) etc. These are broadly categorized as Pro-active and Re-active routing protocols depending upon the type of algorithm used. OLSR is one of the commonly and widely used routing protocol in adhoc networks. It helps in multi-hop communication between the peer to peer nodes connected in adhoc network.

A. OLSR Protocol

OLSR is a proactive routing protocol used by MANETs. RFC 3626 [1] was implemented as OLSR daemon (olsrd) in 2004 [2]. As per RFC 3626, the key concept is Multi-Point Relays (MPRs). Unlike link state routing, where every node transmits broadcast messages; in OLSR only MPRs transmit broadcast messages. Only MPRs generate link state information, hence reducing the number of control messages flooding the network. Thirdly, a MPR may choose to report links between itself and MPR selectors. Learning from experiences of OLSR version 1, RFC 7181 has been issued for latest version i.e. OLSRv2 [3]. It is updated by RFC 7183, 7187, 7188, and 7466.

OLSRv2 is also a table driven, proactive routing protocol which retains the basic mechanisms and algorithms of its predecessor, however, few enhancements have been done in calculation of shortest routes through use of link metric (other than hop count), simplification of messages exchanged and efficient/flexible signaling framework. It is also the further optimizing of classic link state routing protocol and works on concept of MPRs. There are 02 sets of MPRs selected by each router i.e. 'Flooding MPRs' and 'Routing MPRs' used for reduction of flooding and topology, respectively.

1) Implementations of OLSR: Various organizations have carried out implementations and research work on OLSR and is summarized as under. These are primarily based on OLSRv1:

- pyOLSR by INRIA [4]. Initially, pyOLSR was implemented by INRIA as a prototype based on RFC 3626, however, now it is obsolete.
- OOLSR by Hipercom-INRIA[5]. Complete re-implementation in C++ based on RFC 3626 by a Hipercom team at INRIA. It supports Windows and Linux.
- SMOLSR-MOLSR by Hipercom-INRIA[6]. Multicast OLSR (MOLSR) and Simple Multicast OLSR (SMOLSR) are extensions of OOLSR. Their function is to get the control and routing information. The data is forwarded through Multicast Data Forwarding Protocol(MDFP) daemon.
- NOA-OLSR by Niigata University in collaboration with INRIA[7]. No Overhead Auto-configuration OLSR (NOA-OLSR) has been implemented in collaboration with INRIA, based on an Algorithm developed by a professor of Niigata University and INRIA.
- NRL OLSR [8]. It is an implementation based on RFC 3626 with some additional features by Naval Research Laboratories. The code is based on INRIA's OLSR Version 3 protocol draft for the IETF [9]. It has Windows and Linux based distributions.
- QOLSR by LRI [10]. It is a Quality of Service (QoS) extension to OLSR for Linux by Laboratoire de

Recherche en Informatique (LRI), France. Qolyester is RFC 3626 complaint and implemented in C++.

- OLSR by GRC [11]. A Networking Research Group (Grupo de Redes de Computadores (GRC)) at Universitat Politcnica de Valncia (UPV) has ported the OLSR implemented by INRIA to Windows 2000 and Pocket PC.
- OLSR by Unik University [2]. It is the most widely used OLSR implementation and is generally called OLSR Daemon (OLSRd). The olsrd works on Windows (XP and Vista, Windows 7), Linux (i386, arm, alpha, mips, xscale), OS X (powerpc, intel, xscale, iPhone), VxWorks, NetBSD, FreeBSD, OpenBSD, Nokia N900, Google phone (Android) etc [12].

2) OLSRd Plugins – Additional Features: Plugin is an add-on pluggable program fragment enhancing functionality of some program or system. Plugins in context of OLSRd is the supportability to load dynamically loadable library (DLL) for the purpose of performing different functions and to generate or process private package types. In Linux DLL functionality is available in .so files and in Windows as .DLL file extension. The olsrd plugins design has been chosen due to following reasons [2]:

- To add any custom functionality or package, the source code of olsrd is not required to be changed.
- The plugins can be licensed separately as per conditions of user.
- Any language can be used to code the plugins and can be compiled as dynamic library.
- The plugins have backward compatibility.

Various types of plugins for OLSRd are Tiny Web Server (httpinfo), OLSR Node Information Display (txtinfo), Basic Multicast Forwarding (bmf), Securing OLSR Route (secure), Outputs in Dot Format over TCP Socket (dot-draw), Dynamic Announcement of Uplinks (dyn-gw Annoucing), DNS Servers and host names (nameservice), import of external routes from qaugga (qaugga), Performance Graph (pgraph), Distributing P2P Discovery Messages (p2pd), Minimal Example (mini), Tiny Application Server (tas), Detecting OLSRd Freezing (watchdog), optimizing kernel ARP cache from OLSR UDP sniffing (arprefresh), Muticast DNS (mdnsp), Position Update (PUD) etc.

B. Analytical Hierarchical Process (AHP)

AHP is a type of multi-criteria assessment (MCA) technique for analyzing complex decisions. It measures intangibles in relative terms. AHP is a mathematical as well as psychological approach and a structured technique to carry out complex decisions specially applied in group decision making. It was initially studied and researched in the 1970s by Thomas L. Saaty. It has been improved and applied in solving decision problems until now. A good resource on AHP is available at [13]. An AHP process decision making recommends a most suitable choice of alternatives based on user defined criteria. As per this book following steps are involved in an AHP process decision making to recommend a most suitable choice of alternatives based on a defined criteria:

- **Step-1**. The problem is modeled primarily in a hierarchy of three layers as under:
  - The Goal or Main Objective to achieve.
  - The Criteria for evaluation of alternatives.
  - The choice of available alternatives.
- **Step-2**. The elements of hierarchy are pairwise compared based on multiple judgments of each pair of element to establish priorities.
- **Step-3**. The overall priorities of hierarchy are calculated through synthesis of above judgments.
- **Step-4**. Check weather the judgments are consistent and conclude final result based on these judgments.

Various types of tools are available in the market to apply AHP such as BPMSG AHP Online System (AHP-OS), Priority Estimation Tool (PriEst), AHP Solver, MakeItRational, Open Decision Maker, AHP Analyzer, AHP Software, ABC AHP Decision Making Software, easyAHP, AHP.net etc. We in our research have used AHP-OS.

1) BPMSG's AHP-OS: AHP-OS is web based tool developed by Business Performance Management Singapore (BPMSG). It is one of the most latest, updated and easy to use web application/tool (developed in php) available online for Multi Criteria Decision Making (MCDM) based on classical AHP by Saaty. It does not cater other MCDM methods such as Fuzzy AHP, Modified AHP(M-AHP) etc and has peculiar advantages, disadvantages and limitations as associated with each method. It calculates weightings or ratio scales (by paired comparison of criterion) and consistency index based on input by the user (either calculated or subjective opinions). Mathematically it is based on calculation of Eigen value problem. The calculation of Eigen value gives the consistency ratio whereas, the dominant normalized right Eigen vector of the matrix gives the scale ratio. It is based on following features which are available to registered users:

- AHP Projects – A hyperlink to handle complete AHP projects including group decision support.
- AHP Priority Calculator – A hyperlink calculate priorities based on pairwise comparisons.
- AHP Hierarchies – A hyperlink for defining complete set of hierarchies, evaluation of priorities and alternatives.
- AHP Group Session – A hyperlink for participating in AHP group sessions.

2) Application of AHP in Software Selection: A broad review of application of AHP has been presented in [14], [15]. It includes but not limited to selection, evaluation, benefitcost analysis, allocations, planning and development, priority and ranking, decision-making, forecasting in medicines and related fields. Other areas are personal, social, manufacturing sector, political, engineering, education, industry, government, and others which include sports, management etc.

AHP is also applied to selection of software. Simulation Software [16], Multimedia Authorizing Systems (MAS) [17], Project Management Software [18], ETL Software [19], Data Warehouse System for Large and Small Enterprises in Taiwan

[20], Forecasting Software [21] etc are best examples in this regard. These studies motivated us to apply AHP to different variants of OLSR software to select the best one to suit in our adhoc network based collaborative system project.

3) Mathematical Modeling in AHP: A complete and elaborate description of mathematical modeling and application of relevant theorems in AHP is given in [22]. The mathematics used in determining and calculating decision hierarchy and overall result in AHP-OS is given in [23]. The same is summarized in following paragraphs.

- **Scale of Intensity**. The scale of intensity from the 1-9 (denoted by x) is used as an integer for each selection while comparison of paired criteria and alternative. The x is transformed into c (which is used as an element in pairwise comparison matrix) as under:

  Linear scale:
  $$c = x \tag{1}$$

  Logarithmic scale:
  $$c = log_2(x + 1) \tag{2}$$

  Root Square scale:
  $$c = \sqrt{x} \tag{3}$$

  Inverse Linear:
  $$c = \frac{9}{(10 - x)} \tag{4}$$

  Balanced Scale: When w = 0.5, 0.55, 0.6, ..., 0.9.
  $$c = \frac{w}{(1 - w)} \tag{5}$$

  Power Scale:
  $$c = x^2 \tag{6}$$

  Geometric Scale:
  $$c = 2^{x-1} \tag{7}$$

- **Row Geometric Mean Method (RGMM)**. RGMM has been used to calculate the priorities P_i, to input the N x N pairwise comparison of the matrix $A = a_{ij}$. The calculation and normalization is done as under:
  Calculation:
  $$r_i = \exp\left[\frac{1}{N}\sum_{j=1}^{N}\ln(a_{ij})\right] = \Big(\prod_{j=1}^{N} a_{ij}\Big) \tag{8}$$

  Normalization:
  $$p_i = \frac{r_i}{\sum_{j=1}^{N} r_i} \tag{9}$$

- **Consistency Ratio(CR)**. CR is calculated by calculating λ_{max} (the principal eigenvalue) and putting in equation below (calculated by Lonson/ Lamata linear fit):
  $$CR = \frac{\lambda_{max} - N}{2.7699N - 4.3513 - N} \tag{10}$$

II. Defining of Criteria/ Elements of Criterion

A. General Criteria – Mean Opinion Score (MOS) and Relative Opinion Score (ROS)

A general selection criteria based on ROS (as applied in GeoSharing project [24] for selection of an embedded Operating System) where a relative score from 1-5 has been considered (5 being highest score awarded based on observation or judgment of each observer). This type of scoring is relative to one an other (of the projects under consideration) and once carried out by single person can be termed as Relative Opinion Score (ROS). A more relevant ROS can be related as mentioned in Table I. The same type of scoring once conducted through a group of peoples and their average is taken as Mean Opinion Score (MOS).

TABLE I. Relative Opinion Score (ROS) Awards

Observation	Score
Excellent	5
Very Good	4
Good	3
Satisfactory	2
Just Satisfactory	1
Un-Satisfactory	0

B. Elements of Criterion

The selection a software for practical usage depends on multiple factors. However, following factors are considered vital and have been considered as a selection criteria for selection of OLSR software.

- **Stability of software** – Measure of reliability and robustness that it should not crash and it is usable without bugs/ interruptions.
  - OLSRd and NRL-OLSR are quite stable and we assign it ROS of 4. We have experienced crashing of OLSRd and NRL-OLSR very few times.
  - The QOLSR implementation is tested in OPNET simulator. Qolyester is a testbed for QOLSR algorithms,hence, may have bugs. As per available documentation certain bugs are already known. We assign ROS of 3.
  - OLSR's prototype implementation/ porting on Windows 2000 and Pocket PC by GRC-UPV may have inherent bugs. Hence, we have awarded ROS of 2.
  - NOA-OLSR is also a Proof of Concept implementation of OLSR with No Overhead and Auto-Configuration by INRIA, France. Its source is not available for installation and testing. Hence, exact stability can not be ascertained. We have awarded ROS of 1.
  - pyOLSR, OOLSR (developed in C++), SMOLSR and MOLSR (extensions of OOLSR) are also implementations by INRIA France. However, the source code is not available on the referred website. Hence, we assign ROS of 0 to all of these variants with respect to ascertaining the stability.

- **Maintainability by Developers** – Are the developers maintaining there distributions through nightly or stable builds on regular basis?
  - OLSRd is regularly being maintained by www.OLSR.org; latest version is OLSRd 0.9.0.3. OLSRv1 is being used for up-dating and development of OLSR V2. We assign ROS of 5.
  - NRL-OLSR is also being maintained by Naval Research Labs USA. Latest distribution is NRL-OLSRd Ver 7.8.1; last updated on 10 Aug 2007. We assign ROS of 2 to it.
  - QOLSR's most current version is Qolyester-20090302 available online on its website. Since, than it is not updated Hence, we have awarded ROS of 3.
  - NOA-OLSR source is not available for installation and testing. Hence, we have assigned ROS of 0 to it.
  - The source of pyOLSR, OOLSR (developed in C++), SMOLSR and MOLSR (extensions of OOLSR) is also not available and not being updated. Hence, we assigned ROS of 0.
- **Usage** – Is there any developer community which is using the software? User experience of general users and their views are also important factors.
  - As discussed in preceding sub section of Maintainability, same ROS scoring is applied with respect to NRL-OLSR and QOLSR.
  - Considering usage and testing of software by the developer community; we assign ROS of 1 each to NOA-OLSR, OOLSR and PyOLSR.
  - Based on experiences of OLSRd; OLSRv2 is developed and being improved. Moreover, OLSRd has been implemented and used in various adhoc networking projects such as GeoSharing, MANET Manager (SPAN), Byzantium, Commotion, Qual.net etc. Hence, ROS of 5 is awarded for Usage criteria element.
- **Security** – Are security features appropriately addressed in the software and the known vulnerabilities adequately addressed?
  - OLSRd has a security plug-in and we assign ROS of 4.
  - Other variants lacks security feature, hence, we have assigned ROS of 0 to each one of them.
- **Cross-platform** – Does the software support multiple OS platforms such as Windows, Linux, Mac, Android etc.
  - OLSRd is developed for multiple platforms including Windows, Linux, Mac and Android. We award ROS of of 5.
  - The QOLSR is developed for Linux platform. We assign ROS of 2.
  - The NRL-OLSR is also supports multi-platforms and we assign it ROS of 4.
  - NOA-OLSR is also implemented for Linux. We award ROS of 1.
  - pyOLSR is also supports multi-platforms i.e. Windows, POSIX and Linux. We assign ROS of 3.
  - OOLSR is also Windows and Linux based. We assigned ROS of 3.
- **Other Features** – Are multiple features (other than the basic design) provided?
  - OLSRd provides multiple features through plugins as discussed in sub-section above. We award ROS of 4.
  - The QOLSR provides QoS feature. We assign ROS of 1.
  - The NRL-OLSR supports fuzzy-sighted routing and Simplified Multi-cast Forwarding. We assigned ROS of 2.
  - NOA-OLSR supports No Overhead Autoconfiguration; a very important feature in adhoc environment. We award ROS of 1.
  - pyOLSR works with basic OLSR. We assign ROS of 0.
  - OOLSR is improved to provide SMOLSR and MOLSR. We assigned ROS of 2.

C. Summary of ROS

The summary of ROS as per our opinion based on discussion in sub-section II-B are as given in Table II.

TABLE II. RELATIVE OPINION SCORE (ROS) - VARIANTS OF OLSR

Software/ Criteria	OLSRd	pyOLSR	OOLSR	NOA OLSR	NRL OLSR	Q OLSR
Stability	4	0	0	1	4	3
Maintenance	5	0	0	0	2	3
Usage	5	1	1	1	2	3
Security	4	0	0	0	0	0
Cross-platform	5	3	3	1	4	2
Other Features	4	0	2	1	2	1
Total	**27**	**4**	**6**	**4**	**14**	**12**

III. EXPERIMENTATION WITH BPMSG'S AHP-OS

With this definition of basic criteria we move on to our experimentation with BPMSG's AHP-OS. Register with AHP-OS website [22] and login with the provided user name and password. Further steps involved are discussed in ensuing paragraphs.

A. Defining Hierarchy

We defined the hierarchy using the basic node as "Selecting OLSR Software" with node leaf or sub-categories as Stability, Maintenance, Usage, Security, Multi-platform support and Other Features etc. The overall OLSR selection AHP hierarchy along-with requisite criteria and alternatives is as shown in Figure 1.

B. Compare Criteria

Each category of criteria is pairwise compared to find that which criterion has more weight or importance. In our opinion Stability, Maintenance and Security has more importance as compared to Usage, Multi-platform and Other Features criterion of OLSR software. Each pair of criterion was compared in light of following AHP Scale:

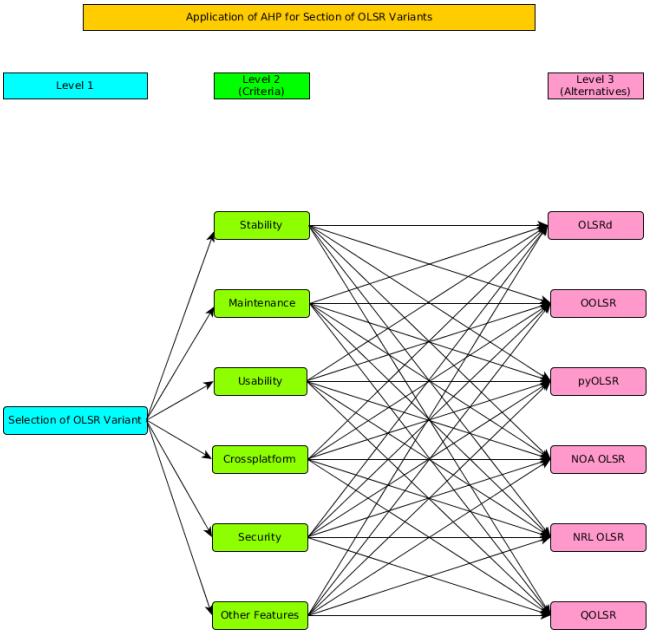


Fig. 1. Overall AHP Hierarchy for Selection of OLSR

Equal Importance as 1, moderate importance as 3, strong importance as 5, very strong importance as 7, Extreme importance as 9, whereas 2, 4, 6, 8 are the values in-between afore stated values.

The selection of pairwise criterion is shown in Figure 2. The overall result of pairwise comparison of each criterion element is shown in Figure 3.

C. Evaluation of Alternatives

On completion of pairwise comparison of criterion, the evaluation of alternatives i.e. OLSR software available in the open source community are also pairwise compared to one another based on subjective opinions or actual measurements. The software also provides group input based decisions as well. The pairwise comparison is made based on the ROS as given in Table II. As an example the pairwise comparison of only one criterion for all the alternatives is shown in Figure 4. Overall status of alternatives is shown in Figure 5. The threadbare analysis of the results recorded in ROS Table II and AHP-OS's Figure 5 reveal that AHP has more granularity with respect to analysis of each criteria element as compared to ROS.

D. Summary of Results

Six OLSR alternatives have been compared in light six criterion elements. Each criterion weighting based on the applicable importance is summarized in graph at Figure 6. Similarly, different variants of OLSR and their overall percentages are summarized as shown in Figure 5. We can clearly see that OLSRd is ranked first, followed by NRL-OLSR as second and QOLSR as third.

Fig. 2. Pairwise Comparison of Criterion Elements

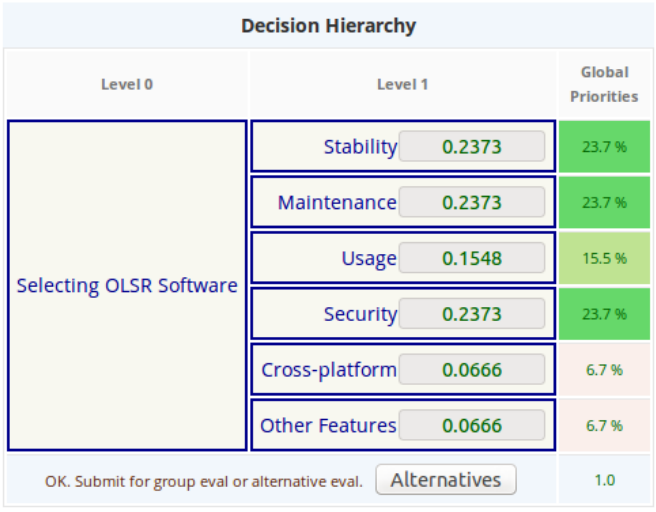


Fig. 3. Pairwise Comparison Results – Criterion Elements

IV. Conclusion and Future Work

OLSRd (by OLSR.org) has been finalized as a better project and a choice for practical implementation for use in a collaborative system being designed in an adhoc networking environment. The empirical study based on quantitative comparison of ROS and AHP criteria elements; ranks OLSRd as First and NRL-OLSR as Second choice for implementation in practical adhoc networks. OLSR2 based on OLSRv2 is considered as the latest implementation based on OLSRd by OLSR.org and can be a better choice for future projects

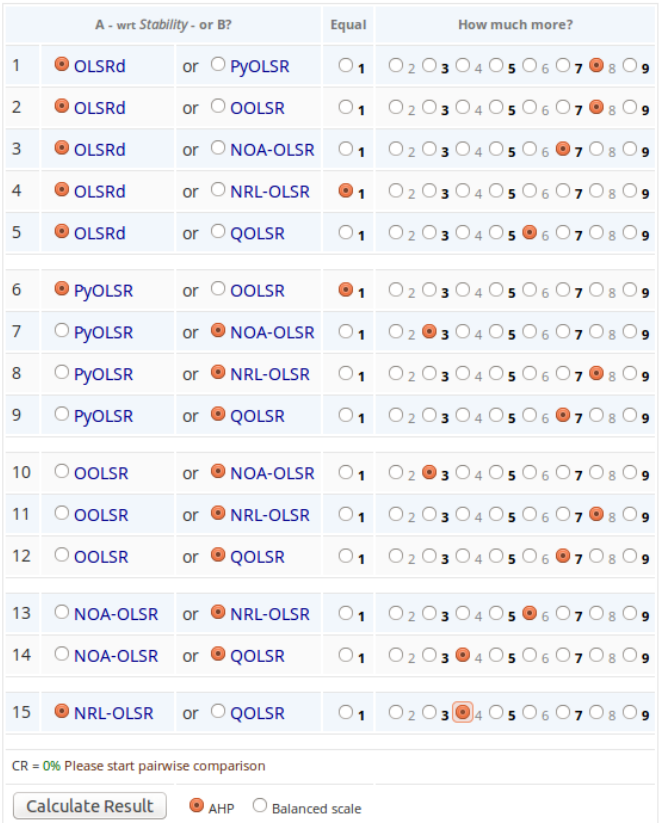


Fig. 4. Pairwise Comparison of Alternatives–Stability

after testing in practical scenarios. Application of AHP-OS decision making helps and provides the objective mathematics to process the inescapable subjective and personal preferences of an individual or a group in making a decision based on various criterion. We have applied and configured a level-2 hierarchy. We intend to apply more relevant sub-criterion to a group of hierarchy such as selection of most suitable adhoc networking project for building a nomadic collaborative information system.

ACKNOWLEDGMENT

We acknowledge the support of NUST-PNEC for carrying out this research by providing necessary funding for procurement of COTS devices. We also thank BPMSG for the free access to on-line AHP-OS Software.

REFERENCES

[1] Network Working Group, "Optimized Link State Routing Protocol (OLSR)." RFC 3626, October 2003.

[2] A. Tnnesen, "Impementing and extending the Optimized Link State Routing Protocol," Master's thesis, Department of Informatics, University of Oslo, August 2004.

[3] Internet Engineering Task Force (IETF), "Optimized Link State Routing Protocol (OLSR) Version 2 (RFC7181)." RFC 7181, April 2014.

[4] Hipercom-INRIA, "pyOLSR." Online Web Page, July 2003. Accessed on 01 August 2015.

[5] Hipercom-INRIA, "OOLSR." Online Web Page, November 2004. Accessed on 01 August 2015.

[6] Hipercom-INRIA, "SMOLSR-MOLSR." Online Web Page, August 2005. Accessed on 01 August 2015.

[7] Hipercom-INRIA and Niigata University, "NOA-OLSR." Online Web Page, 2005. Accessed on 01 August 2015.

[8] Naval Research Laboratories (NRL), "The NRL OLSR Routing Protocol Implementation." Online.

[9] Ron Lee and Joe Macker, "OLSRD HOWTO, Version 1.0." Online, October 2005.

[10] Laboratoire de Recherche en Informatique (LRI), France, "QOLSR." Online Web Page, 2005. Accessed on 01 August 2015.

[11] Grupo de Redes de Computadores (GRC), Networking Research Group-Universitat Politcnica de Valncia (UPV), "OLSR protocol implementation." Online, July 2009. Accessed on 01 August 2015.

[12] Andreas Tnnesen and Thomas Lopatic and Hannes Gredler and Bernd Petrovitsch and Aaron Kaplan and Sven-Ola Tcke, " olsrd - an adhoc wireless mesh routing protocol ." Online Web Page, 2008 (updated). Accessed on 01 August 2015.

[13] T. L. Saaty, *Decision Making for Leaders: The Analytic Hierarchy Process for Decisions in a Complex World*. RWS Publications, 2008.

[14] O. S. Vaidya and S. Kumar, "Analytic hierarchy process: An overview of applications," *Elsevier's European Journal of Operational Research*, pp. 1–29, April 2004.

[15] T. L. Saaty and L. G. Vargas, *Models, Methods, Concepts & Applications of the Analytic Hierarchy Process; International Series in Operations Research & Management Science*, ch. Chapter 2: The Seven Pillars of the Analytic Hierarchy Process, pp. 23–40. Springer Science+Business Media NY, 2012.

[16] S. Erees, E. Kuruolu, and N. Morali, "An application of analytical hierarchy process for simulation software selection in education area," *Frontiers in Science*, vol. 3, pp. 66–70, March 2013.

[17] Vincent S. Lai and Robert P. Trueblood and Bo K. Wong, "Software Selection: A case study of the application of the Analytical Hierarchical Process(AHP) to the selection of a Multimedia Authoring System (MAS)," *Elsevier's Information and Management*, vol. 36, pp. 221–232, January 1999.

[18] B. Kutlu, A. Bozanta, E. Ates, S. Erdogan, O. Gokay, and N. Kan, "Project Management Software Selection Using Analytic Hierarchy Process Method," *International Journal of Applied Science and Technology*, vol. 4, pp. 113–119, November 2014.

[19] M. Hanine, O. Boutkhoum, A. Tiknioune, and T. Agouti, "Application of an integrated multi-criteria decision making AHP-TOPSIS methodology for ETL software selection," *SpringerPlus*, 2016.

[20] H.-Y. Lin and P.-Y. Hsu, "Application of the Analytic Hierarchy Process on Data Warehouse System Selection Decisions for Small and Large Enterprises in Taiwan," *International Journal of The Computer, the Internet and Management*, vol. 15, pp. 73–93, December 2007.

[21] A. Pekin, G. Ozkan, O. Eski, U. Karaarslan, G. Ertek, and K. Kilic, "Application of the Analytic Hierarchy Process (AHP) for Selection of Forecasting Software," *5th InternationalSymposium on Intelligent Manufacturing Systems, Sakarya, Turkey*, 2006.

[22] K. D. Goepel, "BPMSGs AHP Online System." Online, May 2014. Accessed on 02 Sep 15.

[23] K. D. Goepel, "BPMSG AHP Excel Template with multiple Inputs." Online, December 2013.

[24] L. Lamouline and V. Nuttin, "The GeoSharing project: An Openmoko geoposition sharing system," Master's thesis, École Polytechnique de Louvain, Université catholique de Louvain, 2011.

Decision Hierarchy

Level 0	Level 1	Global Priorities	OLSRd	PyOLSR	OOLSR	NOA-OLSR	NRL-OLSR	QOLSR
Selecting OLSR Software	Stability 0.2373	23.7 %	0.0966	0.0081	0.0078	0.014	0.0748	0.036
	Maintenance 0.2373	23.7 %	0.1339	0.0099	0.0103	0.0103	0.0276	0.0453
	Usage 0.1548	15.5 %	0.0897	0.0086	0.0086	0.0086	0.0151	0.0241
	Security 0.2373	23.7 %	0.1384	0.0198	0.0198	0.0198	0.0198	0.0198
	Cross-platform 0.0666	6.7 %	0.0239	0.0089	0.0089	0.0043	0.0164	0.0043
	Other Features 0.0666	6.7 %	0.036	0.003	0.0089	0.0049	0.0089	0.0049
OK. Submit for group eval or alternative eval. Alternatives		1.0	**51.8 %**	**5.8 %**	**6.4 %**	**6.2 %**	**16.3 %**	**13.4 %**

Fig. 5. Overall Result of Pairwise Compared Alternatives and Criterion Elements

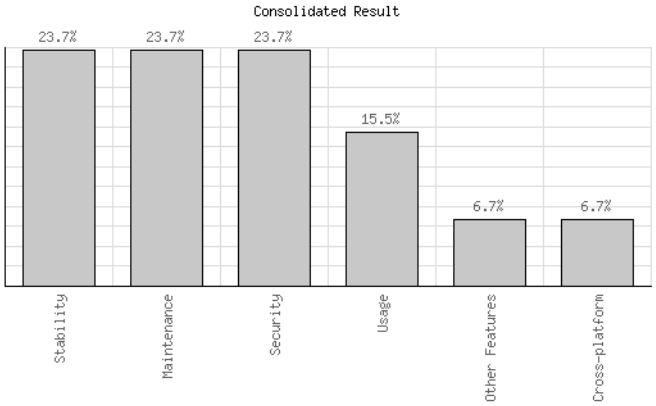


Fig. 6. Summary of Overall Percentage Weight of Criterion Elements

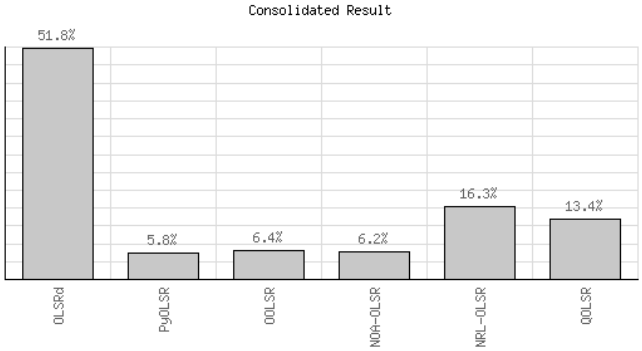


Fig. 7. OLSR Variants Percentages – Participant 1

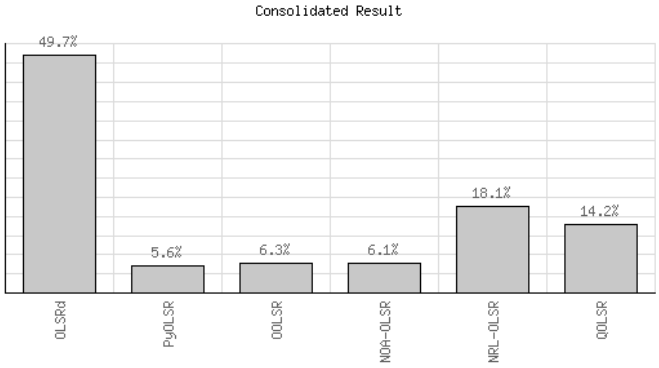


Fig. 8. OLSR Variants Percentages – Participant 2

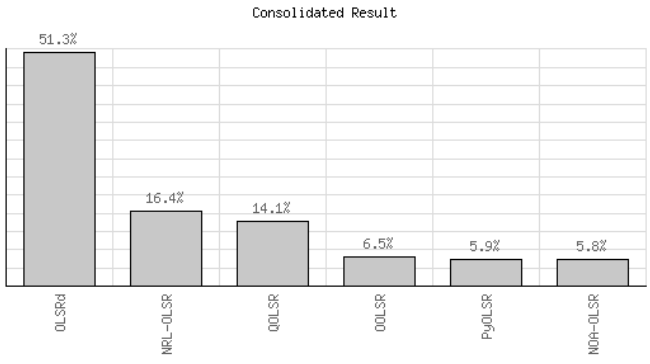


Fig. 9. Overall OLSR Variants Percentages - Average of Both Participants

Fast Approximation for Toeplitz, Tridiagonal, Symmetric and Positive Definite Linear Systems that Grow Over Time

Pedro Mayorga*, Alfonso Estudillo, A. Medina-Santiago, José Vázquez and Fernando Ramos
Faculty of Biomedical Engineering
Universidad Polit´ecnica de Chiapas
Chiapas, M´exico

***Abstract*—Linear systems with tridiagonal structures are very common in problems related not only to engineering, but chemistry, biomedical or finance, for example, real time cubic B-Spline interpolation of ND-images, real time processing of Electrocardiography (ECG) and hand drawing recognition. In those problems which the matrix is positive definite, it is possible to optimize the solution in $\mathcal{O}(n)$ time. This paper describes such systems whose size grows over time and proposes an approximation in $\mathcal{O}(1)$ time of such systems based on a series of previous approximations. In addition, it is described the development of the method and is proved that the proposed solution converges linearly to the optimal. A real-time cubic B-Spline interpolation of an ECG is computed with this proposal, for this application the proposed method shows a global relative error near to 10^{-6} and its computation is faster than traditional methods, as shown in the experiments.**

***Keywords*—real time interpolation; linear convergence; Cholesky decomposition; biomedical data acquisition**

I. Introduction

Several problems in many fields of science are related to linear systems of the form $Ax = b$ [1], [2], [3]. Matrices with special structures such as the Toeplitz matrix arise in many problems, including the solution of ordinary and partial differential equations [4], [5], [6], [2], [7], [8]. In many cases the size of these systems is very large; then, computing the solution in a reasonable amount of time becomes a problem for real time applications [9], [10], [11], [12], [13], [14], [3], [15]. Due to the increasing interest in tridiagonal matrices, it is important to understand its properties and its applications [16], [12], [17], [18], [19], [20], [21], [22], [23], [24].

In general, the system $Ax = b$ is solved in $\mathcal{O}(n^2)$ time [25]; if A is tridiagonal, i.e. $A_{ij} = 0$ for $|i - j| > 1$, there exists fast algorithms to solve it in $\mathcal{O}(n)$ time [16], [11], [2], [19], [3], [25], [24]. Additionally, if A is Toeplitz, symmetric and positive definite as shown in (1), the computation can be optimized [16], [11], [2], [3], for example by using the Cholesky decomposition $A = LL^T$ [11], [3]. Equation (1) describes all of those matrices with constant diagonal and (2) represents the ratio of the matrix coefficients that must lie between -1 and 1.

$$A = \begin{bmatrix} \beta & \alpha & & & & \\ \alpha & \beta & \alpha & & & \\ & \ddots & \ddots & \ddots & & \\ & & \alpha & \beta & \alpha & \\ & & & \alpha & \beta & \end{bmatrix}, \qquad \beta > 2\left|\alpha\right|. \tag{1}$$

$$-1 < \nu \equiv \frac{2\alpha}{\beta} < 1. \tag{2}$$

A. Cholesky decomposition

For a square, symmetric and positive definite matrix A, the Cholesky decomposition computes a matrix L such that $LL^T = A$. The coefficients of L can be computed in $\mathcal{O}(n)$ [3] following the iterative scheme (3):

$$\begin{aligned} L_{11}^2 &= A_{11}, \\ L_{i,i-1} &= A_{i,i-1}/L_{i-1,i-1}, \\ L_{ii}^2 &= A_{ii} - L_{i,i-1}^2, \quad \text{for} \quad i = 2, \ldots, n. \end{aligned} \tag{3}$$

When the matrix A has the structure defined in (1), the coefficients can be computed in a closed form [16], [19], as shown in (4). Those coefficients have the following limits: $\lim_{i\to\infty} L_{ii} = \sqrt{\lambda_1}$ and $\lim_{i\to\infty} L_{i,i-1} = \alpha/\sqrt{\lambda_1}$, due to the fact that $\beta > 2|\alpha|$ and $\lambda_1 > \lambda_2 > 0$.

$$\begin{aligned} L_{ii} &= \sqrt{\frac{\lambda_1^{i+1} - \lambda_2^{i+1}}{\lambda_1^i - \lambda_2^i}}, \\ L_{i,i-1} &= \alpha\sqrt{\frac{\lambda_1^i - \lambda_2^i}{\lambda_1^{i+1} - \lambda_2^{i+1}}}, \quad \text{where} \\ \lambda_1 &= \tfrac{\beta + \sqrt{\beta^2 - 4\alpha^2}}{2}, \ \lambda_2 = \tfrac{\beta - \sqrt{\beta^2 - 4\alpha^2}}{2}. \end{aligned} \tag{4}$$

With this approach the system can be computed in $\mathcal{O}(n)$ and is able to be parallelized [9], [10], [14].

B. Solve $Ax = b$ using Cholesky decomposition

The Cholesky decomposition allows to find the solution x of the linear system $Ax = b$ by first finding u such that $Lu = b$ and, then finding x such that $L^T x = u$. If A is defined as in (1), the Cholesky decomposition can be used to

solve the linear system in $\mathcal{O}(n)$ time. The complete iterative procedure is described as follows:

$$\begin{aligned} u_1 &= \frac{b_1}{L_{11}}, \\ u_i &= \frac{b_i - L_{i,i-1}u_{i-1}}{L_{ii}}, \quad \text{for } i = 2, 3, \ldots, n; \end{aligned} \tag{5}$$

after computing all u_i, the x_i is computed as:

$$\begin{aligned} x_n &= \frac{u_n}{L_{nn}}, \\ x_i &= \frac{u_i - L_{i+1,i}x_{i+1}}{L_{ii}}, \quad \text{for } i = n-1, \ldots, 1. \end{aligned} \tag{6}$$

II. Material and Methods

This section shows the computation of an approximation of the $(n+1) \times (n+1)$ linear system, based on the solution of a similar system of size $n \times n$.

A. A linear system of size $(n+1) \times (n+1)$

Let $Ax = b$ be a linear system, where $A \in \mathbb{R}^{n \times n}$ with structure defined in (1), and $x, b \in \mathbb{R}^n$.

The solution of this kind of systems can be computed with the procedure described in (5) and (6) in $\mathcal{O}(n)$ time. Now, consider an expanded version of the system $Ax = b$ as $A^+x^+ = b^+$, where $A^+ \in \mathbb{R}^{(n+1)\times(n+1)}$ and $x^+, b^+ \in \mathbb{R}^{n+1}$, such that, A^+ has the same structure as (1) and $b_j^+ = b_j$, for $j = 1, 2, \ldots, n$.

The system $A^+x^+ = b^+$ appears in problems that grow over time, such as: real time interpolation of biomedical data (ND Images or ECG), filtering data, handwriting recognition and real time image processing. Solving $A^+x^+ = b^+$ requires additional $\mathcal{O}(n+1)$ time, therefore, solving all systems from size 1 to n is $\mathcal{O}(n^2)$, i.e. this approach is not convenient.

This paper proposes a new method that computes an approximation of $A^+x^+ = b^+$, based on the solution of $Ax = b$ in $\mathcal{O}(1)$ time, moreover, it is demonstrated that the proposed method has linear convergence and it requires only 6 steps to reach a relative error about 10^{-4}. The method is summarized in algorithm 1.

Algorithm 1 Approximation of $A^+x^+ = b^+$ based on $Ax = b$ in $\mathcal{O}(1)$ time

Require: $Ax = b$ of size n with computed solution x.
Require: $A^+x^+ = b^+$ of size $n+1$ s.t. $b_i^+ = b_i$ for $i = 1, \ldots, n$.
Require: an small integer $k > 0$ ▷ typically $k = 6$
1: Set A_{k+1} the matrix (1) of size $k+1$
2: Set $r_1 = b_{n-k+1}^+ - \alpha x_{n-k}$ ▷ update last value
3: Set $r_i = b_{n-k+i}^+$ for $i = 2, \ldots, k+1$
4: Solve $A_{k+1}u = r$
5: Set $x_i^+ = x_i$ for $i = 1, \ldots, n-k$
6: Set $x_{n-k+i}^+ = u_i$ for $i = 1, \ldots, k+1$

The next paragraphs expose the basis of the method as follows:

- Cholesky decomposition used to solve $Ax = b$ and $A^+x^+ = b^+$ shows that intermediate solution u and u^+ share the same values (theorem 1);
- theorem 2 and corollary 1 show that the error between x_i^+ and x_i increases from $i = 1$ to n or decreases from $i = n$ to 1;
- fig. 5 shows that $|x_i^+ - x_i| \propto 10^{-4}$ for $i = 1, \ldots, n-6$. The last point suggests that only is necessary to solve a little system to get an approximation of x^+ as shown in algorithm 1.

Theorem 1. *The first n terms of the partial solution u^+ are identical to the first n terms of the partial solution u, i.e.,*

$$u_j^+ = u_j, \quad \textit{for } j = 1, 2, \ldots, n. \tag{7}$$

Proof: Since A^+ has the same structure as A, then $L_{ij}^+ = L_{ij}$ for $i, j = 1, \ldots, n$; and using $b_i^+ = b_i$ for $i = 1, \ldots, n$:

$$u_1^+ = \frac{b_1^+}{L_{11}^+} = \frac{b_1}{L_{11}} = u_1,$$

and by induction

$$\begin{aligned} u_i^+ &= \frac{b_i^+ - L_{i,i-1}^+ u_{i-1}^+}{L_{ii}^+} = \frac{b_i - L_{i,i-1}u_{i-1}}{L_{ii}} = u_i, \\ &\text{for } i = 2, 3, \ldots, n. \end{aligned}$$

■

Theorem 2. *The absolute error between x_i^+ and x_i is an increasing function of i, moreover*

$$x_i^+ - x_i = (-1)^{n-i+1} \left(\prod_{k=i}^{n} \frac{L_{k+1,k}}{L_{kk}} \right) x_{n+1}^+, \quad i \leq n. \tag{8}$$

Proof: First, the proof of equation (8) is given; second, the proof of equation (8) is shown.

The principle of induction is used to prove (8) as follows: 1) equation (8) is true for $i = n$ (use equation (6)):

$$\begin{aligned} x_n^+ - x_n &= \left(\frac{u_n^+}{L_{nn}^+} - \frac{L_{n+1,n}^+}{L_{nn}^+} x_{n+1}^+ \right) - \left(\frac{u_n}{L_{nn}} \right) \\ &= \left(\frac{u_n}{L_{nn}} - \frac{L_{n+1,n}}{L_{nn}} x_{n+1}^+ \right) - \left(\frac{u_n}{L_{nn}} \right) \\ &= -\frac{L_{n+1,n}}{L_{nn}} x_{n+1}^+; \end{aligned}$$

$\therefore$ (8) is true for $i = n$; 2) assume for a moment that (8) is true for $i = n - j + 1$, given $x_{n-j+1}^+ - x_{n-j+1} = (-1)^j \left(\prod_{k=n-j+1}^{n} \frac{L_{k+1,k}}{L_{kk}} \right) x_{n+1}^+$; 3) the following lines show

the truth for $i = n - j$:

$$\begin{aligned}
&x^+_{n-j} - x_{n-j} \\
&= \frac{u_{n-j} - L_{n-j+1,n-j}x^+_{n-j+1}}{L_{n-j,n-j}} \\
&\quad - \frac{u_{n-j} - L_{n-j+1,n-j}x_{n-j+1}}{L_{n-j,n-j}} \\
&= -\frac{L_{n-j+1,n-j}}{L_{n-j,n-j}}\left(x^+_{n-j+1} - x_{n-j+1}\right) \\
&= -\frac{L_{n-j+1,n-j}}{L_{n-j,n-j}}\left((-1)^j\left(\prod_{k=n-j+1}^{n}\frac{L_{k+1,k}}{L_{k,k}}\right)x^+_{n+1}\right) \\
&= (-1)^{j+1}\left(\prod_{k=n-j}^{n}\frac{L_{k+1,k}}{L_{k,k}}\right)x^+_{n+1}.
\end{aligned}$$

$\therefore$ (8) is true for $i = n - j$.

To prove that $|x^+_i - x_i|$ increases with respect to i the next equation must be satisfied:

$$\frac{|x^+_i - x_i|}{|x^+_{i+1} - x_{i+1}|} = \frac{\left|\prod_{k=i}^{n}\frac{L_{k+1,k}}{L_{kk}}\right||x^+_{n+1}|}{\left|\prod_{k=i+1}^{n}\frac{L_{k+1,k}}{L_{kk}}\right||x^+_{n+1}|} = \left|\frac{L_{i+1,i}}{L_{ii}}\right| < 1.$$

The last expression is true, due to (4) and (2).

■

Corollary 1. *The function* $\gamma_{\nu j} = \left|\prod_{k=n-j+1}^{n}\frac{L_{k+1,k}}{L_{kk}}\right|$ *converges linearly to* 0 *with respect to* j.

Proof: In order to prove that $\gamma_{\nu j}$ converges linearly to 0, it is needed to demonstrate that $\lim_{j\to\infty}\frac{|\gamma_{\nu,j+1}|}{|\gamma_{\nu,j}|} \in (0,1)$ [2]. By using the definition of L in (4):

$$\begin{aligned}
\lim_{j\to\infty}\frac{|\gamma_{\nu,j+1}|}{|\gamma_{\nu,j}|} &= \lim_{j\to\infty}\frac{\left|\prod_{k=n-j}^{n}\frac{\alpha}{\lambda_1}\sqrt{\frac{1-(\lambda_2/\lambda_1)^{n-k}}{1-(\lambda_2/\lambda_1)^{n-k+2}}}\right|}{\left|\prod_{l=n-j+1}^{n}\frac{\alpha}{\lambda_1}\sqrt{\frac{1-(\lambda_2/\lambda_1)^{n-l}}{1-(\lambda_2/\lambda_1)^{n-l+2}}}\right|}, \\
&= \lim_{j\to\infty}\frac{\left|\left(\frac{\alpha}{\lambda_1}\right)^{j+1}\prod_{k=n-j}^{n}\sqrt{\frac{1-(\lambda_2/\lambda_1)^{n-k}}{1-(\lambda_2/\lambda_1)^{n-k+2}}}\right|}{\left|\left(\frac{\alpha}{\lambda_1}\right)^{j}\prod_{l=n-j+1}^{n}\sqrt{\frac{1-(\lambda_2/\lambda_1)^{n-l}}{1-(\lambda_2/\lambda_1)^{n-l+2}}}\right|}, \\
&= \left|\frac{\alpha}{\lambda_1}\right|\lim_{j\to\infty}\left|\sqrt{\frac{1-(\lambda_2/\lambda_1)^{j}}{1-(\lambda_2/\lambda_1)^{j+2}}}\right|;
\end{aligned}$$

the last expression is the product of two values between 0 and 1 $\therefore$ the result is less than 1 and the rate of convergence is linear. ■

The theorem 2 gives us a chance to approximate x^+ with the first $n - j$ terms of x and the last j terms of x^+, i.e. $\tilde{x}^+ \approx [x_1, \ldots, x_{n-j}, x^+_{n-j+1}, \ldots, x^+_n]$. In other words, given the solution of $Ax = b$, an approximation of the expanded system $A^+x^+ = b^+$ is given by the computation of the last j terms of x^+.

Theorem 2 and corollary 1 show that the error between x^+_{n-j} and x_{n-j} decreases linearly with respect to j, as a consequence, the approximation of the expanded system needs few elements and the computation can be done in constant time.

III. Experimental results

In this section the proposed algorithm is tested with three experiments, the first one is a simple example that shows an easy computation of our proposed method, the second one is a real time interpolation of an ECG and a comparison with cubic B-Spline interpolation, and the third one is related to theoretical properties of our method.

A. First numerical example

The following numerical example shows the first validation of the proposed method, by defining two linear systems:

$$\begin{bmatrix}4&1&0&0\\1&4&1&0\\0&1&4&1\\0&0&1&4\end{bmatrix}\begin{bmatrix}x_1\\x_2\\x_3\\x_4\end{bmatrix} = \begin{bmatrix}3\\1\\1\\2\end{bmatrix} \quad \begin{bmatrix}4&1&0&0&0\\1&4&1&0&0\\0&1&4&1&0\\0&0&1&4&1\\0&0&0&1&4\end{bmatrix}\begin{bmatrix}y_1\\y_2\\y_3\\y_4\\y_5\end{bmatrix} = \begin{bmatrix}3\\1\\1\\2\\4\end{bmatrix};$$

with solutions

$$\begin{aligned}
x &= [0.7416, 0.0335, 0.1244, 0.4689], \\
y &= [0.7462, 0.0154, 0.1923, 0.2154, 0.9462], \\
|y - x| &= [0.0045, 0.0181, 0.0679, 0.2535];
\end{aligned}$$

where the two systems share the matrix structure and the same right side results (except for the last value in the second system). Because the second system grows from the first one, it is expected that both solutions x and y are close each other. According to theorem 2, can be deduced that the absolute values $|y_5 - x_5|, \ldots, |y_1 - x_1|$ are in decreasing order and tends to zero as expected.

B. ECG interpolation

Real time interpolation is a common task in data acquisition systems as in an ECG. The cubic B-Spline interpolation is usually applied in computer graphics because its simplicity, quick computation ($\mathcal{O}(\backslash)$) and second order continuity. Even so, in real time applications where new data arrives continuously, there is no sufficiently time to perform the interpolation between samples.

Cubic B-Spline interpolation uses the matrix defined in (4) with $\alpha = 1$ and $\beta = 4$, therefore, it is possible to use the proposed method to perform the approximation of a cubic B-Spline interpolation in real time.

In this example ECG data downloaded from physionet[1] is used:

- Database: MIT-BIH Long-Term ECG Database
- Record: 14046
- Sex: Male

[1] http://physionet.org/cgi-bin/atm/ATM?database=ltdb&record=14046&tdur=3600

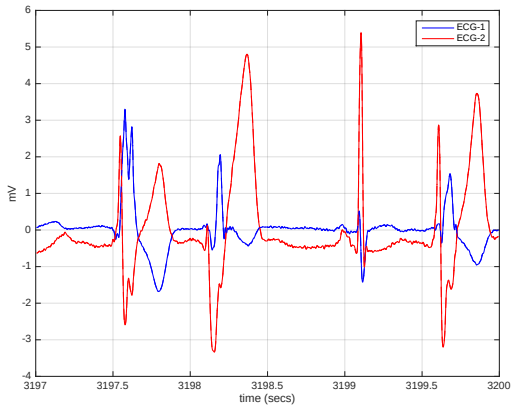


Fig. 1. Cubic B-Spline interpolation of an ECG performed in real time. Average error is about 10^{-6} in both cases.

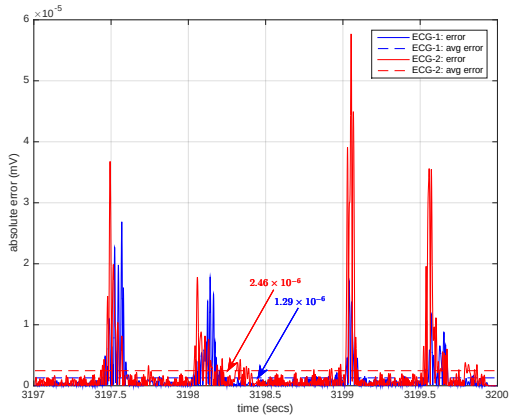


Fig. 2. Cubic B-Spline interpolation of an ECG performed in real time. Average error is about 10^{-6} in both cases.

- Age: 46
- Signals: 2
- Length: 1 hr
- Sampling frequency: 128 Hz
- Sampling interval: 0.0078125 sec
- Samples: 460 800

Because data were collected in 1 hr, it is not a good idea to wait 1 hr to perform cubic interpolation, moreover, when solving the full system every time a data arrives a $\mathcal{O}(n)$ method becomes $\mathcal{O}(n^2)$, i.e. as the system size increases, the computation time increases quadratically. Because our method is $\mathcal{O}(1)$, any approximation is computed in a constant time, and furthermore, for every new data arriving the algorithm becomes $\mathcal{O}(n)$. Computer experiments indicate that solving a system with 460800 elements in size is computed in 0.061 secs using the Cholesky method, this time is greater than the sampling interval, while the proposed method takes 10^{-6} secs independently of the system size and can be used for real time applications.

Interpolation of an ECG is shown in Fig. 1 within an interval from 3197s to 3200s with 383 samples. Cubic B-Spline interpolation was computed and an approximation by using the method described in this paper. In Fig. 1, it is not possible to observe a visual difference between the cubic B-Spline curve and its approximation using the proposed method, this fact is confirmed in Fig. 2 which shows an average error near to 10^{-6}.

Fig. 2 shows the absolute error between the cubic B-Spline interpolation and the proposed method. Cubic B-Spline approximation is near to the exact solution as expected.

Notice that in this case, the error is accumulated through n, and no necessarily is in increasing order.

Now, a comparison of the total computation time used to interpolate ECG samples with the proposed method ($\mathcal{O}(1)$) and the Cholesky method ($\mathcal{O}(n)$) is described. Fig. 3 shows 100 independent experiments of both methods and its averages. A single experiment consists of the total computation time vs system size from $n = 1$ to $n = 16000$. It is possible to observe

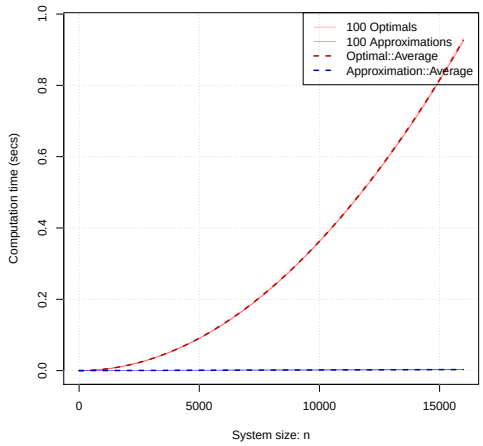


Fig. 3. This plot shows 100 independent experiments to compare the computation time between the Cholesky Method $\mathcal{O}(n)$ in blue lines and our method $\mathcal{O}(1)$ in red lines, in a scenario where the linear system grows from size 1 to 16000 and it is necessary to compute the solution every time the system grows. The cholesky method computes the optimal solution while our method computes and approximation with a relative error of 10^{-6}. In this case the Cholesky method becomes $\sum_{k=1}^{n} \mathcal{O}(k) = \mathcal{O}(n^2)$ and the proposed method becomes $\sum_{k=1}^{n} \mathcal{O}(1) = \mathcal{O}(n)$.

that Cholesky method grows quadratically and the proposed method grows linearly, because $\sum_{k=1}^{n} \mathcal{O}(k) = \mathcal{O}(n^2)$ and $\sum_{k=1}^{n} \mathcal{O}(1) = \mathcal{O}(n)$.

C. Plots and surfaces of the theoretical properties

The matrix A in (4), can be described with a real value because A has one degree of freedom; equation (9) implies $|T(A)| < 1$, $\forall A$ with the structure shown in equation (4).

$$T : \mathbb{R}^{n\times n} \mapsto \mathbb{R} \qquad T(A; \alpha, \beta) = 2\frac{\alpha}{\beta} \tag{9}$$

Fig. 4 shows the log surface of the function proportional to the absolute error between x^+ and x, i.e. $|x^+ - x| \propto \gamma_{\nu j} = \prod_{k=n-j+1}^{n} \left| \frac{L_{k+1,k}}{L_{k,k}} \right|$, in this plot, the j axis represents the number

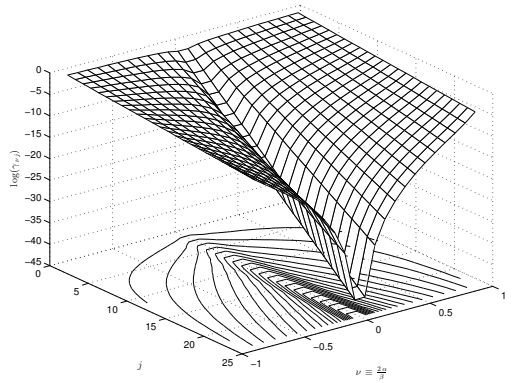


Fig. 4. Log surface of the function $\gamma_{\nu j}$ related to the error of the current and previous solution x^+ and x, respectively; see theorem 2. The parameter j is the index in reverse order that compares elements of x^+ and x, while the parameter $\nu = 2\alpha/\beta$ is related to the matrix structure through the transformation (9). Notice that the error decreases when ν is near to zero, i.e. where the matrix A becomes more diagonally dominant.

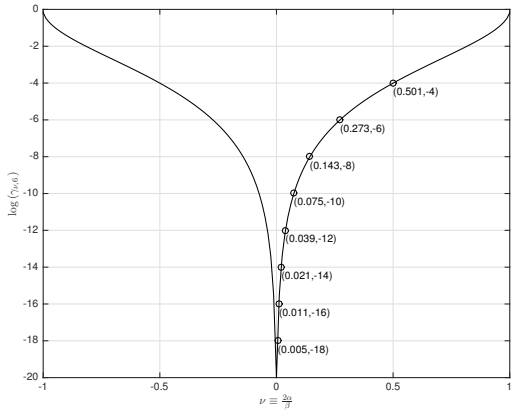


Fig. 5. Slice of the function $\gamma_{\nu j}$ at $j = 6$, see Fig. 4. Marks are located at values where $x^+ - x$ is proportional to 10^{-k}, for $k = 4, 6, \ldots, 18$.

of elements counted from the end, ν is the real value that represents any matrix A through the transformation (9).

Properties related to Fig. 4 are shown bellow:

- $|x_s^+ - x_s| > |x_t^+ - x_t|$ if and only if $s > t$, as seen along the j axis;
- for a fixed ν, the rate of convergence is almost linear;
- for a fixed j, the speed of convergence of $\gamma_{\nu j}$ grows (i.e. $\log \gamma_{\nu j}$ is more negative) when ν goes to zero, indicating that the matrix A is almost diagonal, in this case the computation becomes straightforward.

Fig. 5 shows a slice of the surface in Fig. 4 for $j = 6$, i.e., considering the comparison between the last 6 elements of x^+ and x. Cubic B-Spline and Cubic Bezier interpolation uses a matrix A with $\alpha = 1$ and $\beta = 4$, which gives $\nu = 1/2$ and at this value, the error is about 10^{-4}, see Fig. 5. Table I shows the relative error when the last j terms are computed; it is necessary to compute the last 7 terms of the expanded system for a relative error less than 10^{-4}, and 11 terms are required for a relative error less than 10^{-6}.

TABLE I. NUMBER OF COMPUTATIONS AND ITS RELATIVE ERROR FOR CUBIC B-SPLINE AND CUBIC BEZIER INTERPOLATION, I.E. $\alpha = 1, \beta = 4 \rightarrow \nu = 1/2$

last j terms	Relative error
1	2.6795×10^{-1}
2	7.1797×10^{-2}
3	1.9238×10^{-2}
4	5.1548×10^{-3}
5	1.3812×10^{-3}
6	3.7010×10^{-4}
7	9.9167×10^{-5}
8	2.6572×10^{-5}
9	7.1199×10^{-6}
10	1.9078×10^{-6}
11	5.1118×10^{-7}
12	1.3697×10^{-7}
13	3.6694×10^{-8}
14	9.8069×10^{-9}
15	2.5321×10^{-9}

IV. CONCLUSION

This paper demonstrated that linear systems with special structure (4) (positive definite, tridiagonal, Toeplitz and symmetric) that grow over time, can be approximated in $\mathcal{O}(1)$ time with linear rate of convergence based on a previous solution of a smaller and similar system.

The proofs and properties of the proposed method have been shown. This approach was tested with real time interpolation of an ECG, showing that 1) the average error of the interpolation is about 10^{-6} compared with the exact solution; and 2) the computation time is constant between samples.

The proposed method $\mathcal{O}(1)$ becomes $\mathcal{O}(n)$ while traditional methods of $\mathcal{O}(n)$ becomes $\mathcal{O}(n^2)$ for real time interpolation tasks.

Due to the proposed method properties, this approach can be used in problems where data is generated in real-time environments such as handwriting recognition and interpolation of ND medical images.

ACKNOWLEDGMENT

The authors express their gratitude to the "Universidad Politécnica de Chiapas" for financing this work through the project grant: PRODEP DSA/103.5/14/10628 "Apoyo de fomento a la generación y aplicación innovadora del conocimiento".

REFERENCES

[1] R. Duda, P. Hart, and D. Stork, *Pattern classification*, ser. Pattern Classification and Scene Analysis: Pattern Classification. Wiley, 2001.

[2] J. Nocedal and S. J. Wright, *Numerical Optimization*. Springer, Aug. 2000.

[3] W. H. Press, S. A. Teukolsky, W. T. Vetterling, and B. P. Flannery, *Numerical recipes in C (2nd ed.): the art of scientific computing*. New York, NY, USA: Cambridge University Press, 1992.

[4] J. Dennis and R. Schnabel, *Numerical Methods for Unconstrained Optimization and Nonlinear Equations*, ser. Classics in Applied Mathematics. Society for Industrial and Applied Mathematics, 1996.

[5] F. Diele and L. Lopez, "The use of the factorization of five-diagonal matrices by tridiagonal Toeplitz matrices," *Applied Mathematics Letters*, vol. 11, no. 3, pp. 61–69, 1998.

[6] D. Fischer, G. Golub, O. Hald, C. Leiva, and O. Widlund, "On fourier-toeplitz methods for separable elliptic problems," *Mathematics of Computation*, vol. 28, no. 126, pp. 349–368, 1974.

[7] G. D. Smith, *Numerical Solution of Partial Differential Equations*, 2nd ed. Clarendon Press: Oxford, 1978.

[8] J. Stewart, *Calculus: Early Transcendentals*, ser. Textbooks Available with Cengage YouBook Series. Cengage Learning, 2010.

[9] O. Axelsson, *Iterative Solution Methods*. Cambridge University Press, 1996.

[10] A. B. Boudewijn, E. Bell, and B. R. Haverkort, "Serial and parallel out-of-core solution of linear systems arising from generalised stochastic petri nets," pp. 22–26, 2001.

[11] J. Dennis and R. Schnabel, *Numerical Methods for Unconstrained Optimization and Nonlinear Equations*, ser. Classics in Applied Mathematics. Society for Industrial and Applied Mathematics, 1996.

[12] P. Dewilde and K. Diepold, "Large-Scale Linear Computations with Dedicated Real-Time Architectures," in *Advances in Real-Time Systems*, S. Chakraborty and J. Eberspächer, Eds. Springer Berlin Heidelberg, pp. 41–81, 2012.

[13] C. Gerald, *Applied numerical analysis*. Addison-Wesley, 1980.

[14] M. James, G. Smith, and J. Wolford, *Applied numerical methods for digital computation*, 1st ed., ser. Applied Numerical Methods for Digital Computation. Harper & Row, 1985.

[15] H.-T. Yau and J.-B. Wang, "Fast bezier interpolator with real-time lookahead function for high-accuracy machining," *International Journal of Machine Tools and Manufacture*, vol. 47, no. 10, pp. 1518 – 1529, 2007.

[16] S. Chandrasekaran, M. Gu, X. Sun, J. Xia, and J. Zhu, "A Superfast Algorithm for Toeplitz Systems of Linear Equations," *SIAM Journal on Matrix Analysis and Applications*, vol. 29, no. 4, pp. 1247–1266, 2008.

[17] J. Feng, Y. Li, Y. Wang, and M. Chen, "Design of a real-time adaptive nurbs interpolator with axis acceleration limit," *The International Journal of Advanced Manufacturing Technology*, vol. 48, no. 1-4, pp. 227–241, 2010.

[18] M.-T. Lin, M.-S. Tsai, and H.-T. Yau, "Development of a dynamics-based {NURBS} interpolator with real-time look-ahead algorithm," *International Journal of Machine Tools and Manufacture*, vol. 47, no. 15, pp. 2246 – 2262, 2007.

[19] S. Noschese, L. Pasquini, and L. Reichel, "Tridiagonal Toeplitz matrices: properties and novel applications," *Numerical Linear Algebra with Applications*, vol. 20, no. 2, pp. 302–326, 2013.

[20] L. Piegl and W. Tiller, *The NURBS Book*, ser. Monographs in Visual Communication. U.S. Government Printing Office, 1997.

[21] M.-S. Tsai, H.-W. Nien, and H.-T. Yau, "Development of a real-time look-ahead interpolation methodology with spline-fitting technique for high-speed machining," *The International Journal of Advanced Manufacturing Technology*, vol. 47, no. 5-8, pp. 621–638, 2010.

[22] Y. Wang, D. Yang, and Y. Liu, "A real-time look-ahead interpolation algorithm based on akima curve fitting," *International Journal of Machine Tools and Manufacture*, vol. 85, no. 0, pp. 122 – 130, 2014.

[23] H. Zhao, L. Zhu, and H. Ding, "A real-time look-ahead interpolation methodology with curvature-continuous b-spline transition scheme for {CNC} machining of short line segments," *International Journal of Machine Tools and Manufacture*, vol. 65, no. 0, pp. 88 – 98, 2013.

[24] W. Zheng, P. Bo, Y. Liu, and W. Wang, "Fast B-spline curve fitting by L-BFGS," *Comput. Aided Geom. Des.*, vol. 29, no. 7, pp. 448–462, Oct. 2012.

[25] G. Strang, *Linear Algebra and Its Applications*. Brooks Cole, Feb. 1988.

A Multi-Agent Framework for Data Extraction,Transformation and Loading in Data Warehouse

Ramzan Talib*, Muhammad Kashif Hanif†, Fakeeha Fatima‡, and Shaeela Ayesha§
Department of Computer Science,
Government College University, Faisalabad, Pakistan

***Abstract*—The rapid growth in size of data sets poses challenge to extract and analyze information in timely manner for better prediction and decision making. Data warehouse is the solution for strategic decision making. Data warehouse serves as a repository to store historical and current data. Extraction, Transformation and Loading (ETL) process gather data from different sources and integrate it into data warehouse. This paper proposes a multi-agent framework that enhance the efficiency of ETL process. Agents perform specific task assigned to them. The identification of errors at different stages of ETL process become easy. This was difficult and time consuming in traditional ETL process. Multi-agent framework identify data sources, extract, integrate, transform, and load data into data warehouse. A monitoring agent remains active during this process and generate alerts if there is issue at any stage.**

***Keywords*—Data Warehouse; Extraction; Loading; Multi-Agent; Operational Data; Transformation**

I. Introduction

In this digital era, data is being generated by different sources at all times. Data can reside on different computers and servers. It is challenging task for organizations to manage and analyze huge volume of data to achieve their goals [1]. Combination of historical and current data is essential to get strategic information. Data warehouse provide the input for strategic decision making. Data warehouse takes data from different sources, process and store in a common repository [2]. Data warehouse is a subject oriented, integrated, time variant and nonvolatile collection of data in support of taking management decision [2]. Data warehouse provide accurate, efficient, and complete view of an organization's operational data to solve complex queries [3].

The most important component of the data warehouse is ETL process. ETL process extract and integrate data from diverse homogeneous and heterogeneous sources. Data sources may contain inconsistent data that can produce incorrect and misleading results. The purpose of ETL process is to extract data from data sources, transform into structured format, and load into target data warehouse [4], [5].

Figure 1 shows the generic architecture of data warehouse. This architecture consists of data source, mapping of ETL process, storage area and analysis layers. The purpose of analysis layer is to mine data for future predication [6], [7].

The data store format in data warehouse is different from operational data sources. Operational systems are essential for day to day operations of any organization. In Online Transaction Processing (OLTP) system, records are stored in flat files using different applications, tools, data formats, and data representational methods [8]. Efficient and effective data extraction methods are essential to make better and reliable strategic decision [9]. A process that converts operational data into analytical data stores in a controlled, secure and suitable format is needed [4]. The ETL process provides data cleaning consistently and reliably with high performance. Different tools can be used at this phase like talend, pentaho, oracle warehouse builder, Microsoft integration services, open text integration services, IBM cognos managers, information builder, SQL server integration tool, and SSIS packages etc. [10], [11].

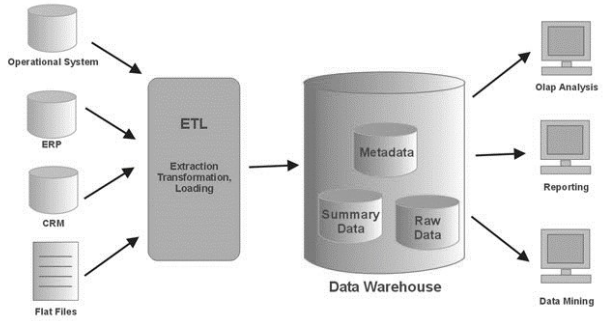


Fig. 1. Architecture of data warehouse [3]

There is need to extract more meaningful, relevant and appropriate data to take reliable, efficient and effective decisions. For this, a multi-agent based framework in ETL process is proposed. The proposed process will make the data extraction, transformation and loading process fast, efficient, and flexible. In this way, efficiency and effectiveness to manage and extraction of data is increased [12].

The remainder of this paper is organized in different sections. In section II, related work is discussed. Section III presents ETL process in data warehouse. Section IV provides a multi-agent framework in ETL process. Section V concludes the paper with future research perspectives.

II. Review of Literature

[5] presented a framework for extraction, transforming and loading data into data warehouse. They have shown extraction is most important phase in ETL process. They also discussed the issues for extracting, transforming and loading

data and their effect on the decision making process. Futher, various quality metrics and ad hoc approaches that enhance the performance of ETL process were proposed. [1] presented different modeling techniques that optimize and enhance the ETL process. Different techniques to design ETL process were discussed in field of academic. These techniques were based on open source and commercial tools. They concluded ETL process is very expensive regarding its cost, time and establishment of data warehouse.

[4] proposed a multi-agent based framework to establish a data warehouse structure for information technology infrastructure library. The use of agents in data warehouse optimizes and enhances the working. ITIL is relatively a complex and large data warehouse infrastructure which manages all IT related fields. By following standards with multi-agent technology help to manage the continuous updation, improve functionality and reduce the chances of risks. [13] presented a multi-agent system used at the data pre-processing stage in e-wedding project. The use of Multi-Agent System (MAS) at earlier stage improve responsiveness and efficiency of the system. A multi-agent system based on Java Agent Development Framework (JADE) is used to cope these issues raised at data pre-processing stage, i.e., handle missing values during data extraction process. JADE support different states of the agents as: agent communication, protocol, behavior, and ontology.

[14] presented novel methods to handle complex data consolidation through multi-agent system in data warehouse. The proposed approach based on more flexible and evaluative architecture in which one can easily add, remove and modify services according to the need. By applying multi-agent based prototype, the integration process done by using UML classes. Two agents the data agent and wrapper agent were used to model the complex data in UML classes. The XML creator agent mapped the UML classes into XML document. [15] discussed how to improve Extraction Transformation and Loading (ETL) process in data warehouse system of higher education system. They presented ETL architecture for HEIS and discussed various issues which arise in development and maintenance of the data model.

[16] discussed the use of agent technology in data warehouse. They have proposed Intelligent Data Warehouse (IDW) model for data extraction, processing and information retrieval optimization. In this model, data is integrated from different sources efficiently. Moreover, data collection process is improved which reduces the extraction time. It incorporates functions that are adaptable and flexible to access the data across the enterprises. [17] presented the workflow process for the refreshment of data. They concluded data updation process should be performed in extraction, transformation and loading phases.

III. Extraction, Transformation and Loading Process

The complexity and disparity of sources is growing with the increased usage of information systems. Data warehouse provides central repository that enable organizations to store all historical data at one place. ETL is a most important phase to extract, clean and transform data in data warehouse. Figure 2 shows data flow in ETL process. ETL process has extraction, transformation, and loading steps (Figure 3).

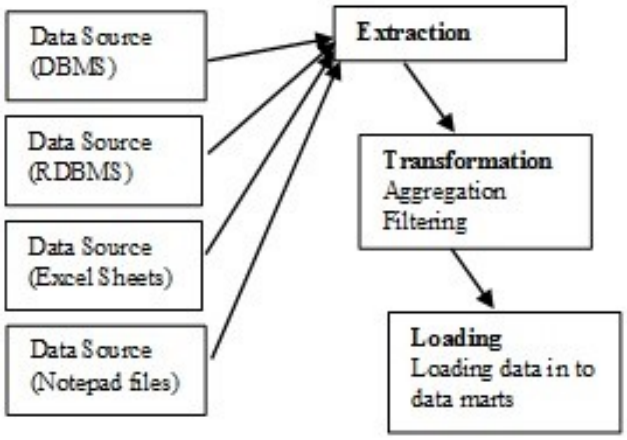


Fig. 2. ETL data flow diagram [5]

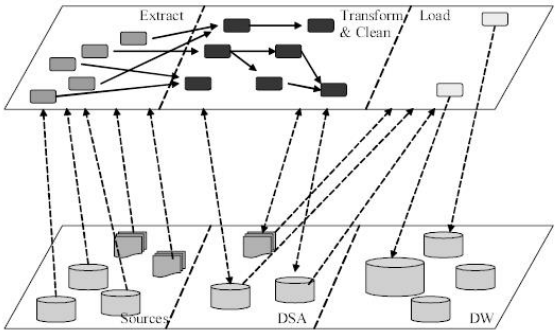


Fig. 3. ETL process [6]

A. Extraction

The first phase of ETL process is data extraction. This phase integrate data from various homogeneous and heterogeneous sources. The source systems can have different data format according to their operational needs [18]. Data validation rules are applied on the operational data according to domain. Validation rules identify whether the data extracted from the sources has the correct values [19]. Following constraints are checked at this stage [10].

- content and meta data
- data object attributes
- extraction mode and protocols to capture data
- monitor the extraction process

B. Transformation

Transformation is a most crucial phase. At the staging area, different mapping functions are performed on the extracted data to remove dirty values, duplications, inconsistencies, and naming conventions [17]. Manipulation operations like cleansing, filtering, enriching, aggregating, sorting, generating surrogate keys, and granularity level are determined to map

the external data source to data warehouse [18], [20]. Set of rules to translate coded values and to derive new values are applied to clean and transmit the data. At this phase, schema and instance level mapping are performed to standardize the data. In addition, data validation and data accuracy constrain are performed [21].

C. Loading

This is a final phase in ETL process. Extracted and transformed data are loaded into targeted data warehouse [14]. Data loading in the data warehouse has its own technical challenges. A major problem is difference between new and existing data at loading time. This step make sure data is converted into targeted data structure of data warehouse rather than source data structures. Moreover, various schema and instance joining and aggregation functions are performed at this phase [15].

IV. A Multi-Agent Framework in ETL Process

Intelligent agents are used now a days in every field of life to solve complex problem by distributing the work. Agents are a software programs that take the autonomous action in different states to attain design objectives. According to [1], responsive, proactive, independent, object oriented and social are important characteristics of agents. In multi-agent based system, agents work collectively and each agent performs specific tasks according to the role assigned [14], [15].

The addition of agents at the data extraction level minimizes the chance of error, increases efficiency and reliability (Figure 4). Moreover, the extraction, transformation and loading time is reduced [22]. Agents invoke messages when any problem arises. An alert is generated for missing value or irrelevant data [10], [16].

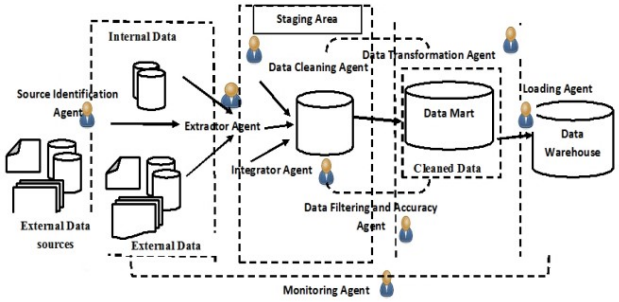


Fig. 4. The Multi-Agents framework for ETL

There is no specific standard and structures for operational data. Discrepancies arise in data formats due to changing characteristics of data [19]. Multi-agents in ETL process helps to reduce the chances of occurrence of errors. Each agent is assigned a specific role by following the standards regarding the semantic and format of data [23]. In this case study, agent based ETL process is analyzed that helps to make the extraction, transformation and loading process efficient, effective, and reliable [21], [24]. Agents are organized into three groups in multi-agent framework.

- Extractor and Integrator Multi-Agent Group
- Transformator and Loader Multi-Agent Group
- Management and Control Agent

A. Extractor and Integrator Multi-Agent Group

Agents in this group extract data from different excel files, flat files, MS access and SQL databases [8]. Agents coordinate with other agents in this group to extract complete, concise, and reliable data (Figure 5). Agents in this group are assigned following roles.

- **Source Identifier Agent (SIA)** identify the data extraction sources.
- **Extractor Agent (EA)** establishes a link with the sources system and extracts data.
- **Data Cleaning Agent (DCA)** is concerned with identifying and eliminating contradictions and inconsistencies. DCA removes duplicate, missing and irrelevant values from data. It is also responsible for the customization and integration of the information from multiple sources [25].
- **Integrator Agent (IA)** integrates and mounts the extracted data in the Data Staging Area (DSA). The extracted records are loaded into data warehouse staging area (Figure 5). At this stage, DCA removes all discrepancies of spelling error, invalid or wrong records to improve the quality and reliability of the data [26].

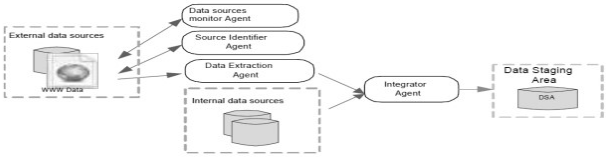


Fig. 5. The Extractor and Integrator Multi-Agent Group

B. Transformator and Loader Multi-Agent Group (TLM)

Transformator and Loader Multi-Agent Group is responsible for data validation, accuracy, consistency, schema and instance related conversion according to semantic rules [25]. TLM follows a work-flow sequence to execute all data transformation in a reliable and efficient way. This multi-agent group is consists of the following agents:

- **Data Validator Agent (DVA)** checks and matches all records of the fact and dimension tables to ensure integrity constraints.
- **Data Filtered and Accuracy Agent (DFAA)** make sure record contains appropriate values and mapped according to data warehouse structure [27].
- **Loader Agent (LA)** is responsible for loading record from logical schema into repository mapped schema. The role of LA is to ensure efficiency and consistency to improve the performance of data warehouse operations and reduce the loading time.

C. Management and Control Agent (MCA)

The purpose of MCA in ETL process is to monitor all the activities of agents. MCA ensure agents are doing work properly and according to the sequence. It also establishes a coordination among agents to enhance the functionality and performance.

V. Conclusion

A multi-agent based ETL framework provide an efficient mechanism to extract, transform and load data in data warehouse. ETL process extract data from homogeneous, heterogeneous, or distributed sources and map in the format according to targeted data warehouse. There exist different methods and tools to enhance the efficiency of ETL process. In this work, an agent based framework is proposed. In proposed framework, agents work collectively to perform tasks according to the roles assigned. The system contains EIM, TLM, and MCA groups of agents to reduce the extraction time and optimize the performance. Research can be carried to design a common model for the meta data of ETL process. Moreover, the implementation of the agent based scenario for analysis purpose in different fields of life can be done.

References

[1] A. Bologa, R. Bologa *et al.*, "Business intelligence using software agents," *Database Systems Journal*, vol. 2, no. 4, pp. 31–42, 2011.

[2] W. H. Inmon, *Building the data warehouse*. John wiley & sons, 2005.

[3] Z. El Akkaoui, E. Zimànyi, J.-N. Mazón, and J. Trujillo, "A model-driven framework for ETL process development," in *Proceedings of the ACM 14th international workshop on Data Warehousing and OLAP*. ACM, 2011, pp. 45–52.

[4] P. Christen, "A survey of indexing techniques for scalable record linkage and deduplication," *IEEE transactions on knowledge and data engineering*, vol. 24, no. 9, pp. 1537–1555, 2012.

[5] P. Balaji and D. Srinivasan, "An introduction to multi-agent systems," in *Innovations in multi-agent systems and applications-1*. Springer, 2010, pp. 1–27.

[6] M. Arif and G. Mujtaba, "A survey: Data warehouse architecture," *International Journal of Hybrid Information Technology*, vol. 8, no. 5, pp. 349–356, 2015.

[7] M. Golfarelli and S. Rizzi, *Data warehouse design: Modern principles and methodologies*. McGraw-Hill, Inc., 2009.

[8] V. Gour, S. Sarangdevot, G. S. Tanwar, and A. Sharma, "Improve performance of extract, transform and load (ETL) in data warehouse," *International Journal on Computer Science and Engineering*, vol. 2, no. 3, pp. 786–789, 2010.

[9] S. H. A. El-Sappagh, A. M. A. Hendawi, and A. H. El Bastawissy, "A proposed model for data warehouse ETL processes," *Journal of King Saud University-Computer and Information Sciences*, vol. 23, no. 2, pp. 91–104, 2011.

[10] A. KABIRI and D. CHIADMI, "Survey on ETL processes." *Journal of Theoretical and Applied Information Technology*, vol. 54, no. 2, 2013.

[11] M. Mrunalini, T. S. Kumar, and K. R. Kanth, "Simulating secure data extraction in extraction transformation loading (ETL) processes," in *Computer Modeling and Simulation, 2009. EMS'09. Third UKSim European Symposium on*. IEEE, 2009, pp. 142–147.

[12] M. Singh and S. Jain, "Transformation rules for decomposing heterogeneous data into triples," *Journal of King Saud University-Computer and Information Sciences*, vol. 27, no. 2, pp. 181–192, 2015.

[13] N. Kolsi, A. Abdellatif, and K. Ghedira, "Data warehouse access using multi-agent system," *Distributed and Parallel Databases*, vol. 25, no. 1-2, pp. 29–45, 2009.

[14] A. J. Morais, E. Oliveira, and A. M. Jorge, "A multi-agent recommender system," in *Distributed Computing and Artificial Intelligence*. Springer, 2012, pp. 281–288.

[15] K. Kularbphettong, G. Clayton, and P. Meesad, "A hybrid system based on multi-agent system in the data preprocessing stage," *arXiv preprint arXiv:1003.1792*, 2010.

[16] O. Boussaïd, F. Bentayeb, and J. Darmont, "An mas-based ETL approach for complex data," *arXiv preprint arXiv:0809.2686*, 2008.

[17] M. Wooldridge, *An introduction to multiagent systems*. John Wiley & Sons, 2009.

[18] M. M. Al-Debei, "Data warehouse as a backbone for business intelligence: Issues and challenges," *European Journal of Economics, Finance and Administrative Sciences*, vol. 33, 2011.

[19] G. Cong, W. Fan, F. Geerts, X. Jia, and S. Ma, "Improving data quality: Consistency and accuracy," in *Proceedings of the 33rd international conference on Very large data bases*. VLDB Endowment, 2007, pp. 315–326.

[20] F. Fatima, M. Javed, F. Amjad, and U. G. Khan, "An approach to enhance quality of the rad model using agents," *The International Journal of Science and Technology*, vol. 5, pp. 2002–2010, 2014.

[21] I. Mekterović, L. Brkić, and M. Baranović, "Improving the ETL process and maintenance of higher education information system data warehouse," *WSEAS transactions on computers*, vol. 8, no. 10, pp. 1681–1690, 2009.

[22] L. Muñoz, J.-N. Mazón, J. Pardillo, and J. Trujillo, "Modelling ETL processes of data warehouses with uml activity diagrams," in *OTM Confederated International Conferences" On the Move to Meaningful Internet Systems"*. Springer, 2008, pp. 44–53.

[23] R. Gill and J. Singh, "Enactment of medium and small scale enterprise ETL (masseetl)-an open source tool," *International Journal of Computer Science and Information Technologies*, vol. 6, no. 1, pp. 141–147, 2015.

[24] C. A. Moturi and A. Emurugat, "Prototyping an academic data warehouse: Case for a public university in kenya," *British Journal of Applied Science & Technology*, vol. 8, no. 6, 2015.

[25] A. Abello, O. Romero, T. B. Pedersen, R. Berlanga, V. Nebot, M. J. Aramburu, and A. Simitsis, "Using semantic web technologies for exploratory OLAP: a survey," *IEEE Transactions on Knowledge and Data Engineering*, vol. 27, no. 2, pp. 571–588, 2015.

[26] K. Sivaganesh, P. Srinivasu, and S. C. Satapathy, "Optimization of ETL work flow in data warehouse," *International Journal on Computer Science and Engineering*, vol. 4, no. 9, p. 1579, 2012.

[27] M. Jarke, M. Lenzerini, Y. Vassiliou, and P. Vassiliadis, *Fundamentals of data warehouses*. Springer Science & Business Media, 2013.

Polynomial based Channel Estimation Technique with Sliding Window for *M*-QAM Systems

O. O. Ogundile*, M. O. Oloyede†, F. A. Aina‡ and S. S. Oyewobi§
*Department of Physics and Telecommunications, Tai Solarin University of Education Ijagun, Ogun State, Nigeria
†Department of Information and Communication Science, University of Ilorin, Ilorin, Kwara State, Nigeria
‡Department of Telecommunication Science, University of Ilorin, Ilorin, Kwara State, Nigeria
§Department of Telecommunication Engineering, Federal University of Technology, Minna, Nigeria

***Abstract*—Pilot Symbol Assisted Modulation (PSAM) channel estimation techniques over Rayleigh fading channels have been analysed in recent years. Fluctuations in the Rayleigh fading channel gain degrades the performance of any modulation scheme. This paper develops and analyses a PSAM Polynomial interpolation technique based on Least Square (*LS*) approximations to estimate the Channel State Information (CSI) for *M*-ary Quadrature Amplitude Modulation (*M*-QAM) over flat Rayleigh fading channels. A Sliding window approach with pilot symbol adjustment is employed in order to minimize the computational time complexity of the estimation technique. The channel estimation performance, and its computational delay and time complexity is verified for different Doppler frequencies (f_d), frame lengths (L), and Polynomial orders (*P-orders*). Simulation results show that the Cubic Polynomial interpolation gives superior Symbol Error Rate (SER) performance than the Quadratic Polynomial interpolation and higher *P-orders*, and the performance of the Polynomial estimation techniques degrade with increase in the *P-orders*.**

***Keywords*—Channel estimation; Doppler frequency; frame length; interpolation; polynomial order**

I. Introduction

Channel estimation assuming a Rayleigh fading channel with Additive White Gaussian Noise (AWGN) is an essential aspect of wireless communication. Most especially if Orthogonal Frequency Division Multiplexing (OFDM) or *M*-ary Quadrature Amplitude Modulation (*M*-QAM) is adopted as the modulation scheme in conveying information between the transmitter and receiver. Rayleigh channel gain variations result in degradation in the performance of the modulation scheme, more significantly at low Signal-to-Noise Ratio (SNR) [1]. Therefore, channel estimation is required to obtain a rough estimate of the Channel State Information (CSI), and compensate for such channel errors.

Pilot Symbol Assisted Modulation (PSAM) channel estimation techniques have been proposed in literature using Gaussian, Wiener, Sinc, Linear interpolations, and many more interpolation techniques to analyse and compensate for the effect of fading errors in order to improve the performance of the modulation scheme [2]–[8]. The major focus has been on the trade-off between the computational time complexity and gain performance of these channel estimation techniques over Rayleigh fading channels.

This paper proposes a Polynomial interpolation channel estimation technique based on Least Square (*LS*) approximations for *M*-QAM assuming a flat Rayleigh fading channel. The Polynomial estimation technique offers a balance between the performance and implementation time complexity in estimating the CSI. Simulations are provided to verify the optimal performance of the interpolation technique for different Polynomial orders (*P-orders*), while increasing the Doppler frequency f_d, and frame length size L. Result is also shown comparing the developed Polynomial interpolation technique with some existing channel estimation techniques found in [4], [7] and [9]. The simulation results are shown assuming rectangular 16-QAM, although the estimation technique can be extended to other *M*-QAM schemes as well.

The remainder of this paper is structured as follows. Section II gives a brief explanation of the system model and notations. In section III, the PSAM Polynomial interpolation technique is developed and analysed for different *P-orders*. Discussions and result comparisons are presented in Section IV. Finally, the paper is summarised and concluded in Section V.

II. System Model

Fig. 1 shows a pilot symbol assisted modulation system model. The input data is mapped to rectangular 16-QAM complex data symbols with the real and imaginary components taken from the set $(\pm 1c, \pm 3c, ... \pm (m-1)c)$, where m is of the form $M{=}m^2$ [10]–[12]. Pilot symbols are periodically inserted to the transmitted 16-QAM complex data symbols. The transmitted QAM symbols are divided into frames of size L, with each frame starting with a known symbol called the pilot symbol, and subsequently L-1 data symbols. The frame structure is transmitted over a flat Rayleigh fading channel with Doppler frequency f_d.

After down-converting and matched filtering, the faded received signal corrupted with AWGN in the current *p-th* frame is defined as:

$$\bar{r}_p^{\,l} = \bar{E}_p^l \bar{t}_p^{\,l} + N_p, \qquad l = 1, ..., L-1, \tag{1}$$

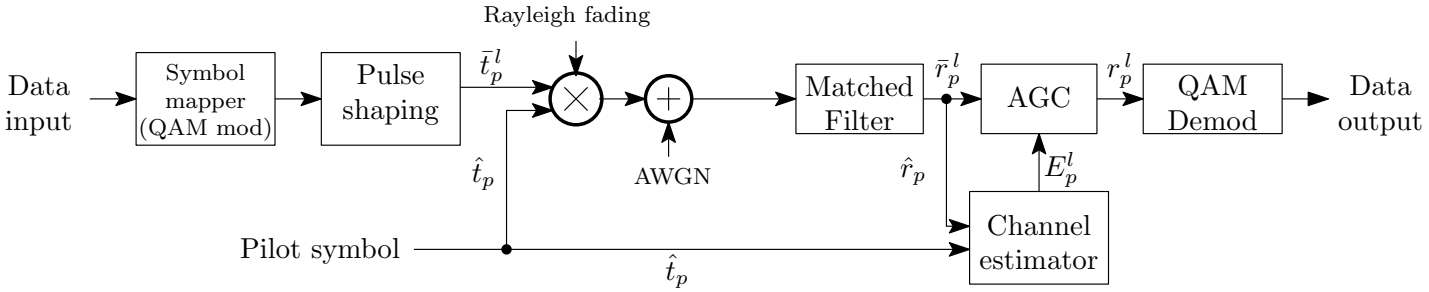


Fig. 1: PSAM system model.

where N_p is the complex AWGN with variance $N_o/2$, $\bar{t}_p^l$ is the transmitted QAM symbols, and $\bar{E}_p^l$ is the complex zero mean Gaussian variables denoting the fading distortion at the data symbol positions.

Similarly, the faded received pilot symbol corrupted with AWGN in the current *p-th* frame is defined as:

$$\hat{r}_p = \hat{E}_p \hat{t}_p + N_p, \tag{2}$$

where $\hat{E}_p$ is the complex zero mean Gaussian variables denoting the fading distortion at the pilot symbol positions, $\hat{t}_p$ and $\hat{r}_p$ are the transmitted and received pilot symbols respectively. Thus, the noisy fading estimates at the *p-th* pilot symbol positions is obtained as:

$$E_p = \hat{E}_p + \frac{N_p}{\hat{t}_p}. \tag{3}$$

The fading estimates E_p^l in the current *p-th* frame at the data symbol position is obtained by *interpolating* the pilot symbol fading estimates E_p. The received data symbol $\bar{r}_p^l$ is scaled (dividing the received data symbol $\bar{r}_p^l$ by the estimated fading distortion E_p^l) to compensate for the channel error in a process called Automatic Gain Control (AGC). The scaled received data symbol r_p^l is demodulated to obtain a copy of the transmitted data.

III. PSAM Polynomial Interpolation

The Polynomial interpolation channel estimation technique maintains a balance between the performance and implementation time complexity in estimating the CSI. The fading distortion in the current *p-th* frame at the data symbol position is estimated by interpolating the current *p-th* frame pilot symbol fading estimate E_p and the nearest pilot symbol fading estimates E_{p+i} to the *p-th* frame. The number of the nearest pilot symbol estimates used for the interpolation process depends on the *P-order*.

Consider the complex data and pilot symbols frame structure at the output of the matched filter as shown in Fig. 2. Given that $(E_p, E_{p+1}, \ldots, E_{p+i})$ are the nearest pilot symbol fading estimates to the *l-th* data symbol and $(K_p, K_{p+1}, \ldots, K_{p+i})$ are the known positions of the pilot symbols in the received frame structure of Fig. 2. The fading distortion is computed in the current *p-th* frame as:

$$E_p^l = \sum_{l=1}^{L-1} \gamma_1 (l)^0 + \gamma_2 (l)^1 + \ldots + \gamma_{i+1} (l)^i, \tag{4}$$

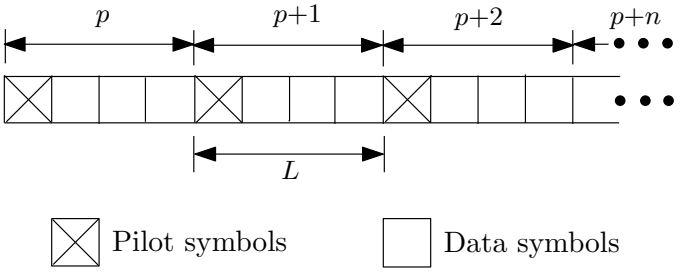


Fig. 2: Data and pilot symbols frame structure.

where i is the *P-order*, and $(\gamma_1, \gamma_2, ..., \gamma_{i+1})$ are the Polynomial interpolation coefficients which is define as:

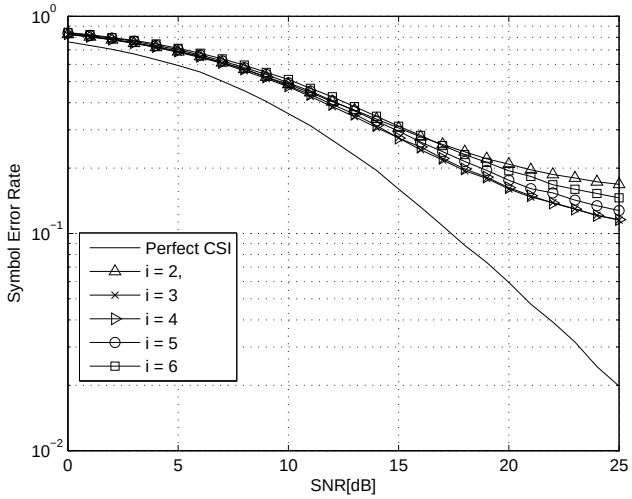
$$\begin{bmatrix} E_p \\ E_{p+1} \\ \vdots \\ E_{p+i} \end{bmatrix} = \begin{bmatrix} 1 & K_p & \ldots & K_p^i \\ 1 & K_{p+1} & \ldots & K_{p+1}^i \\ \vdots & \vdots & \ddots & \vdots \\ 1 & K_{p+i} & \ldots & K_{p+i}^i \end{bmatrix} \begin{bmatrix} \gamma_1 \\ \gamma_2 \\ \vdots \\ \gamma_{i+1} \end{bmatrix}. \tag{5}$$

This estimation technique experiences delay with increase in the frame length size L but the computational time complexity remains the same. Thus, the computational time complexity of the Polynomial interpolation technique is independent of L. However, as the *P-order* increases, the estimation technique experiences delay and the computational time complexity increases at a linear rate, $(O(n)) \rightarrow (O(n^i))$. In order to reduce the computational time complexity of this estimation technique from $(O(n^i)) \rightarrow (O(n^{i-1}))$, a sliding window approach is assumed in computing the Polynomial interpolation coefficients.

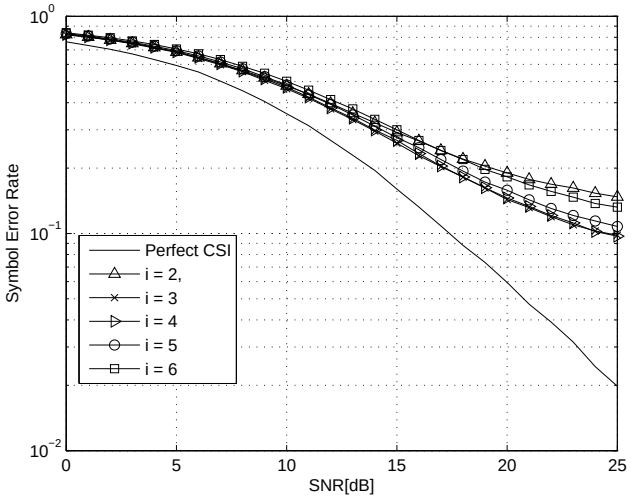
The nearest pilot symbol positions are fixed as $(K_1, K_{1+L}, \ldots, K_{1+iL})$, while the fading estimates at the pilot symbol positions are shifted from E_p to E_{p+1} when computing the Polynomial interpolation coefficients from frame p to $p+n$. Therefore, Eqn. 5 is modified as:

$$\begin{bmatrix} E_p \\ E_{p+1} \\ \vdots \\ E_{p+i} \end{bmatrix} = \begin{bmatrix} 1 & 1 & \ldots & 1 \\ 1 & K_{1+L} & \ldots & K_{1+L}^i \\ \vdots & \vdots & \ddots & \vdots \\ 1 & K_{1+iL} & \ldots & K_{1+iL}^i \end{bmatrix} \begin{bmatrix} \gamma_1 \\ \gamma_2 \\ \vdots \\ \gamma_{i+1} \end{bmatrix}. \tag{6}$$

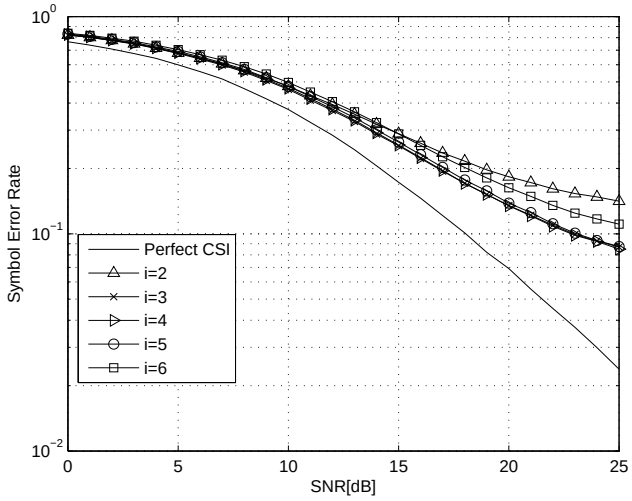
The transmitted frame structure starts with a pilot symbol and subsequently L-1 data symbols, so K_1 is always equal



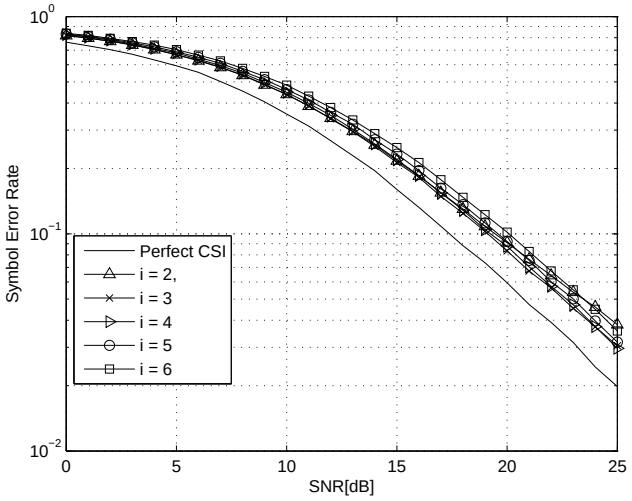
(a) Frame length L=4 and Doppler frequency f_d=200Hz



(b) Frame length L=6 and Doppler frequency f_d=150Hz



(c) Frame length L=8 and Doppler frequency f_d=100Hz



(d) Frame length L=10 and Doppler frequency f_d=50Hz

Fig. 3: SER performance comparison for different *P-orders i* and Doppler frequencies f_d assuming rectangular 16-QAM over flat Rayleigh fading channels.

to one (K_1=1). Applying Eqns. 6 and 4, computing the fading estimates at the data symbol positions in the current *p-th* frame is straightforward.

IV. Results and Discussion

The performance analysis of the Polynomial based channel estimator is carried out assuming a 16-QAM scheme over flat Rayleigh fading channel. The estimator is applied for different values of Doppler frequency f_d; from 50Hz which denotes a practical slow-varying flat Rayleigh fading channel to a fast-varying flat Rayleigh fading channel (100Hz-200Hz). The performance of the polynomial channel estimator is also tested for increase in *P-order* and frame length size L as shown in Fig. 3, in order to verify the optimal *P-order* for real time channel estimations. Note that a different frame length size L is used as the Doppler frequency is increased from 50Hz to 200HZ. This is because the performance of most channel estimator plummets as L increase. Thus, in order to achieve a reasonable Symbol Error Rate (SER) performance for higher Doppler frequency, a smaller L is used in the simulation set-up as f_d increases.

The Polynomial estimator provides a practical computational time complexity of ($O(n^{i-1})$) to the overall system model of Fig. 1 for large input data size. This implies that the computational time complexity of the Polynomial estimator increases with increase in the *P-order*.The Polynomial estimator provides a practical computational time complexity of ($O(n^{i-1})$) to the overall system model of Fig. 1 for large input data size. This implies that the computational time complexity of the Polynomial estimator increases with increase in the *P-order*.The Polynomial estimator provides a practical computational time complexity of ($O(n^{i-1})$) to the overall system model of Fig. 1 for large input data size. This implies that the computational time complexity of the Polynomial estimator

increases with increase in the *P-order*.The Polynomial estimator provides a practical computational time complexity of ($O(n^{i-1})$) to the overall system model of Fig. 1 for large input data size. This implies that the computational time complexity of the Polynomial estimator increases with increase in the *P-order*. Thus, the Quadratic polynomial interpolator (i=2) offers the minimum computational time complexity compared to all other *P-orders* ($2 \leq i \leq \infty$), and it experiences the lowest computational delay.

However, as shown in Fig. 3, the Cubic polynomial interpolator (i=3) yields better SER performance compared to the Quadratic interpolator, and higher Polynomial interpolator orders. It is also observed from Fig. 3 that as the *P-order* increases from i=3 → i=6 → i=∞, the SER performance of the Polynomial interpolator degrades, and the computational delay and time complexity increases as well. The Cubic interpolator gives a SER gain of $+1.5dB$ over the Quadratic interpolator at SNR $>15dB$ as shown in Fig. 3(a)-(c), and it offers moderate computational time complexity. Hence, the Cubic polynomial estimator provides a balance between the performance of the channel estimation technique and its implementation time complexity. The Cubic estimator therefore produces the "optimal" performance among all Polynomial estimation technique of the form of Eqns. 6 and 4 over flat Rayleigh fading channels.

The optimal performance of the Polynomial channel estimation technique (i = 3) is compared with two existing channel estimation techniques in literature. Fig 4. shows the SER performance of the Cubic estimator with the Sinc interpolator proposed be Kim *et al*. [4], and the Linear interpolator analysed in [7] and [9]. The estimators are applied assuming a fast-varying flat Rayleigh fading channel (Doppler frequency $f_d = 100Hz$), and for frame length sizes $L = 4$ and $L = 6$. The Cubic interpolator yields a significant SER performance of $+1.5dB$ over the Linear, and Sinc channel estimators at SNR $>15dB$. Also, Fig. 4 shows that the Cubic interpolator at frame length $L = 6$ outperforms the Sinc, and Linear interpolators at frame length $L = 4$. Although, the proposed Polynomial estimation technique has the price of increased computational time complexity compared to the Sinc, and Linear estimators, it is a useful channel estimation technique and can be implemented for real time systems over Rayleigh fading channels.

V. Conclusion

A pilot symbol assisted modulation Polynomial channel estimation technique has been developed and analysed assuming rectangular 16-QAM over a flat Rayleigh fading channel. The performance of the Polynomial channel estimation technique is tested for different *P-orders*, while increasing the Doppler frequency f_d and frame length L in order to find out the practical optimal performance of the proposed estimator. The estimator technique attains its optimal performance at *P-order* i=3, whereby it maintains a balance between the performance of the channel estimation technique and its computational time complexity. The optimal performance (i=3) of the Polynomial channel estimation technique gives better SER performance than the Sinc, and Linear estimators over fast flat Rayleigh fading channel. It is noted that the developed Polynomial channel interpolator evinces considerably higher computational time complexity compared to the Sinc, and Linear interpolators; however, it offers significant SER performance gain and can be implemented for real time systems. Results provided in Fig. 3 and 4 affirm these analysis.

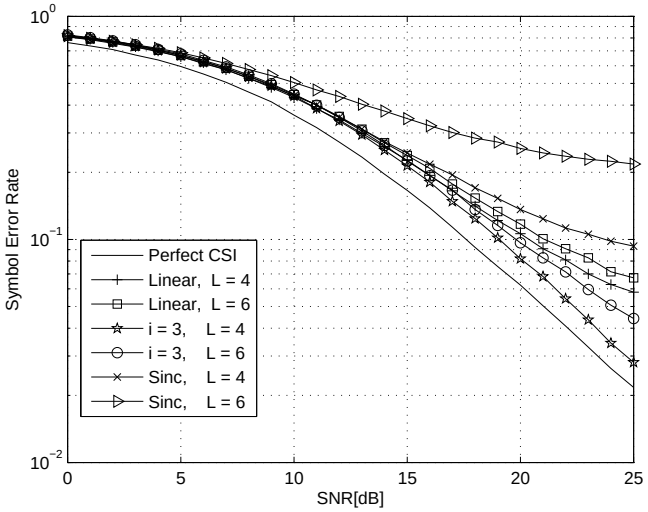


Fig. 4: SER performance comparison for the Cubic, Sinc, and Linear estimators assuming rectangular 16-QAM over a fast flat Rayleigh fading channel. Doppler frequency, $f_d = 100Hz$.

References

[1] Theodore S. Rappaport, *Wireless communications: principles and practice*. Prentice Hall PTR, 2nd edition, September 2002.

[2] J. Cavers, "An Analysis of Pilot Symbol Assisted Modulation for Rayleigh Fading Channels." *IEEE, Transaction on Vehicular Technology*, vol. 40, no. 4, pp. 686–693, November 1991.

[3] S. Sampei and T. Sunaga, "Rayleigh Fading Compensation for QAM in Land Mobile Radio Communications." *IEEE ,Trans. Veh. Technol.*, vol. 42, pp. 137–146, May 1993.

[4] Y. Kim, G. Jeong, Y. Bang, H. Park, and S. Choi, "New Rayleigh Fading Channel Estimator Based on PSAM Channel Sounding Techniques." *IEEE*, vol. 3, pp. 1518–1520, 1997.

[5] X. Tang, M. Alouini, and A. Goldsmith, "Effect of Channel Estimation Error on *M*-QAM BER Performance in Rayleigh Fading." *IEEE,Trans. Commun.*, vol. 47, pp. 1856–1864, December 1999.

[6] S. Coleri, M. Ergen, A. Puri, and A. Bahai, "Channel Estimation Techniques Based on Pilot Arrangement in OFDM Systems," *IEEE Transactions on Broadcasting*, vol. 48, no. 3, September 2002.

[7] M. Benjillahi and L. Szczeciński, "Low Complexity Channel Estimation with Pilot Symbol Assisted Modulation." *IEEE, Signal Processing and Its Applications*, vol. 2, pp. 471–474, August 2005.

[8] Y. Shen and Ed Martinez, "Channer Estimation in OFDM Systems," *freescale semiconductor, AN3059*, January 2006.

[9] A. Mämmelä and V. Kaasila, "Smoothing and Interpolation in a Pilot-Symbol Assisted Diversity System." *International Journal of Wireless Information Networks*, vol. 4, no. 3, 1997.

[10] M. K. Simon and J. G. Smith, "Carrier Synchronization and Detection of QASK Signal Sets," *IEEE, Transactions on Communication*, vol. 22, no. 2, pp. 98–106, February 1974.

[11] O. Ogundile and D. Versfeld, "Improved reliability information for rectangular 16-QAM over flat rayleigh fading channels," in *Computational Science and Engineering (CSE), 2014 IEEE 17th International Conference on*, Dec 2014, pp. 345–349.

[12] ——, "Improved reliability information for OFDM systems on time-varying frequency-selective fading channels," in *Wireless Telecommunications Symposium (WTS), 2015*, April 2015, pp. 1–7.

Synergies of Advanced Technologies and Role of VANET in Logistics and Transportation

Kishwer Abdul Khaliq
Department of Production Engineering,
IGS,University of Bremen, Germany

Amir Qayyum
CoReNeT,
Capital University of Science and
Technology (CUST), Islamabad, Pakistan

Jürgen Pannek
Department of Production Engineering
University of Bremen, Germany

***Abstract*—In Intelligent Transport Systems (ITS), Vehicular Ad-hoc Network (VANET) is one of key wireless technologies, which helps in managing road safety, traffic efficiency, fleet, logistics and transportation. The objective of this paper is to give an overview of the implication of different technologies and placement of VANET in transportation and specifically in logistics. We provide researchers with an overview of considered technologies in logistics scenarios and the current projects regarding VANET for safety and non-safety applications. We additionally discuss current and potential domains in logistics in which new applications can improve efficiency by use of new and existing technologies.**

***Keywords*—VANET; IEEE802.11p; Logistics; Vehicular Ad-hoc Network; Transportation; Technology role**

I. Introduction

In project of the European commission (EC) named "Mobility and Transport", Intelligent Transport Systems (ITS) [1] is one of the transport topics which deals with the traffic management, safety and efficiency among many target transport mode. To get improvement in transportation system, it applies information and communication technologies like computers, electronics, satellites and sensors. These technological possibilities require us to rethink design, implementation and deployment of existing technologies in different transport modes like road, air, water, and rail to provide new services for passengers and freight transport [2]. Therefore, the first goal of ITS is to manage transport systems and the second goal is to render the transport network more safe. In case of the traffic congestion, it aims to reduce both the traffic and also its impact on the environment. Hence, efficiency is the third goal of ITS. To achieve the goal of traffic management, safety, and efficiency, a number of communication technologies are involved in ITS, e.g., ITS-G5 [3], Wi-Fi [4], 3G [5], LTE [6] etc. to create innovative solutions. The linkage between the EC project, ITS and technologies is shown in Figure 1. Several technologies have been deployed to maintain and promote ITS. Logistics companies focus on the flexibility and efficiency to save time and labor cost. Thus, the improvement of the transportation system and communicating interfaces of these companies contributes to achieve the goals [7].

Vehicular Ad-hoc Network is the one of the challenging domains in wireless networks and has unique features. It does not only offer efficient traffic management, logistics and transportation, navigation, and road safety applications, but also regards for online gaming and infotainment applications [8].

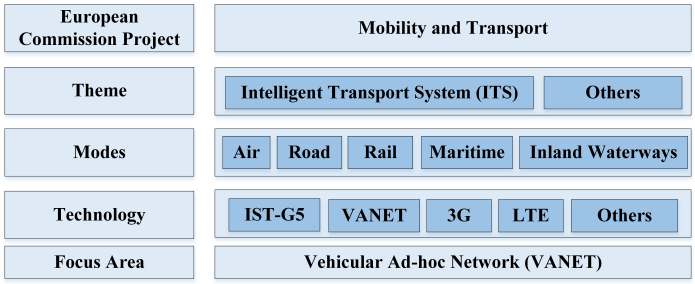


Fig. 1: Intelligent Transport System (ITS) and Underlying Technologies

Logistics include a number of activities that ensure the timely availability of the right product to the customer and these activities create a bridge between production and consumption [9]. Consequently, it links the production and market locations. Two parameters can have great impact on the distance between production plant to the location market or supplier unit, i.e. availability of the product to deliver, and time efficiency in delivery of it. To get market benefits, delivery of the right product to the right consumer on right time is equally important for all companies. These two requirements have changed the procedure of production and delivery. To cope with this challenge, the sub-system of logistic requires efficiency and automation. Figure 2 shows the logistics process, where each stage has specific requirements with respect to the next level. At each level different technologies are in use to get the required benefits. Regarding raw material, the efficient collection of it can boost the production of the desired product. At the production units, synchronization of production steps autonomously and information management about the product specification according to demand are the important factors. Latest technologies like robotics, WiFi, multi-agent system and others help to increase productivity and improve management systems. Market share increases with the best distribution of product in the market and simple accessibility for the customer. Online technologies play an important role not only for the product advertisement, but also for purchasing.

Each company promotes its product as a best product in the market. A new product progresses through a sequence of stages called product life cycle from introduction to growth, maturity, and decline. The success of the production cycle depends upon type of product, knowledge of production, people and

knowledge of people's requirements. In order to increase the value of the product, different services are added to the product by converting raw-materials into the customized products. Therefore, the production cycle includes many values into it. For example form-value is added to the product by converting the raw-materials into finished product during production and manufacturing, place-value is provided through transportation by moving the finished products to the needed location, and time-value is provided through storage and inventory controls to ensure the availability of the products when needed. Finally, possession-value is added to the product through marketing and sales. In this whole procedure, place and time-values are the key logistics functions.

To meet the challenge of product-delivery from the production-unit to the market, efficiency in transportation is required. Efficiency in terms of time may vary because it depends on the type of product, mode of transport and location of the need [10]–[12]. The development of technology like automobiles, electronic devices, home appliances require different place and time-values than the production of food items. The food-item exhibits a short life span and requires delivery to market when it is fresh. In a competitive market, the latest technologies are used to shorten the process of production, and enhance storage and inventory for the quick distribution, monitoring and possible re-routing. Sourcing from

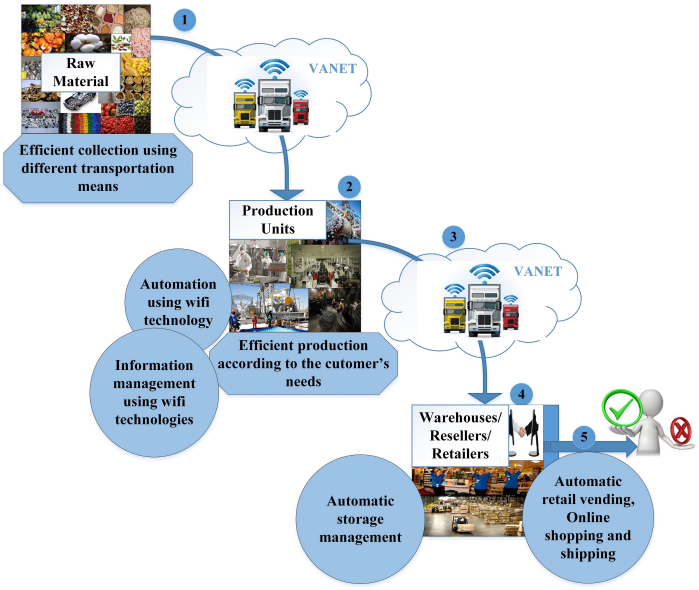


Fig. 2: Possible Placement of Technologies in Logistics and Transportation

the raw material to the finished products and the respective distribution involves many tiers in the supply chain flow. In practice, supply chain integration is the set screw and also active research area to improve the supply chain performance. Basically, this integration involves two kinds of flows. The first flow involves the physical steps that need to be carried out, while the second flow complements the first flow (logistics) with respective information. Previous studies [13]–[15] addressed these two flows by merging information and logistics. The research studies also showed that with the help of technology transference [16]–[18] and new technologies integration e.g., RFID [19], robotics [20], [21], WiFi and communication technologies [22]–[25] etc., the process of logistics and transportation has become more flexible. The objective of this paper is to pinpoint the advantages of existing technologies and importance of VANET technology in logistics and transportation with respect to the previously mentioned issues, and to discuss use of multiple technologies together to a get solution for complex processes. The use of technologies aims to simplify this process by reducing time of product life cycle with value added services and reducing delay while delivering products.

The rest of the paper is organized as follows. Section II gives the state of the art. It reviews the technology role in logistics and transportation, explains different technologies in context of different sub-problems and challenges in the scenario of logistics and transportation, and also discusses VANET current projects in the research area. Section III discusses different scenarios where one or more technologies can add benefits to companies. It also describes applications of different technologies for logistics and transportation. Section IV concludes and explains possible future work.

II. Advanced Technologies in Logistics

The efficient production of customized products and their supply are the key to success for many companies. Failure leads to loss of revenue, decline in level of services, reputation and market share. Recent developments in the market e.g. increase in market competition along number of products with short life cycle and product proliferation, have created a scenario where the customer's demands are unpredictable. Thus, the ability to appropriately respond to the market has become a major asset for many companies, and a motivation for improving their logistics systems [26].

The last few decades, the business environment has changed due to advances in information technologies in extracting, manufacturing and servicing industries. In addition to these, the positive growth of knowledge industry [27] [28] has raised productivity by generating more worker autonomy or greater managerial control. This change is particularly visible in the European Union. Many small and medium size companies have logistic management at their high priority to gain competitive benefits [29], [30]. From point of origin to the point of destination, logistics includes planning, implementing, controlling, transportation of goods, services and related information [31]. The reverse logistics system planning [32] is beneficial for home appliances and proposed a mixed integer programming model to determine the optimal configuration. This model used return rates to determine the numbers, storage locations and plants and showed the benefits of sharing facilities in recycling electrical appliances and computers [33]. To obtain an efficient and flexible system start-to-end, different technologies can be used for sub-tasks. Figure 4 divides the logistics tasks and sub-functions. In each sub-function, integration of particular technology offers added values for the efficient execution and best end results. Each technology plays a vital role to solve sub-functions of logistics to earn revenue.

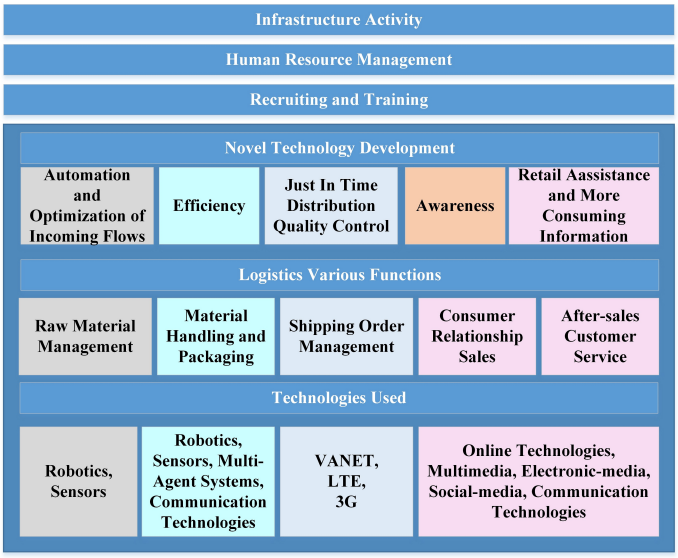


Fig. 3: Logistics Process with Integration of Technologies

A. Multi-Agent Systems

Logistics and supply chain systems consist of number of sub-system that can communicate with one another autonomously to maximize utility. Considering their complex relationship and decision making processes of these sub-system, a multi-agent system modeling is a suitable approach. It is used for the convenient modeling of these sub-systems [34] to improve efficiency of supply chain, solve dynamic logistic management, food supply chain management, fleet management, mass customisation issues and make profit through e-commerce.

To improve the efficiency of the food supply chain, the authors of [29] introduced an agent technology model and resulting food supply chain includes the new feature of intelligence, which allows to optimize performance of the system. They described two ways for optimization: firstly, they reviewed intelligent agents applications, analyzed and compared the existing technologies, then critically reviewed the integration of agent technology in the supply chain management. Secondly, they explained the multi-agent system and its mechanism of optimization to solve many tasks like inventory issues, bullwhip effect, communication problem or adverse risk sharing. Furthermore, it offers a capability to purchase and sale while in transit. In [21], the authors discussed the convenience of automatic systems for loading and unloading with the increase of the number of parcels. This automation can offer benefits to logistics companies by reducing the cost of labor and consumption of time.

To solve the problem of dynamic logistic process management, a real-time knowledge-based framework has proposed in [35] with the use of RFID-multi-agents. The system is capable of real-time process management" which has functionality to identify current process status, perform checking/reasoning, and support to staff members by providing knowledge about the process while handling logistics activity problems. It included an application case study of Eastern Worldwide Company Limited to reveal the performance of operations and resources utilization significantly.

The authors in paper [36] discuss multi-agent system for the case study of e-commerce. Here, for the design of the logistics, delivery of products, their storage and transportation, a third party logistics (3PL) takes responsibility in a supply chain. A 3PL vendor used multi-agent system to build up a private logistics service unit and virtual private logistics subsystem (VPL). By this way, they integrated own logistics business process with the supply chain members. In the supply chain, the logistics and its information flow are seamlessly connected with the financing and trading flows. The 3PL vendor would keep the stability of its own business process while providing customized services to the supply chain members.

The resource allocation problem is usually solved, when system has real time information of all the orders and resources in advance and it does not affect the process of scheduling. On the basis of real time analysis, it is concluded that intelligent applications help to cover mentioned problem in domain of transport system including traffic issues. These intelligent systems include multi-agent simulation platform for traffic modeling, decision support systems for letter transportation, logistics planning, sea freight transportation, vehicle dispatching, scheduling for railway and truck transportation and others [37]–[42] [43].

In [41], authors defined mass customization as a transition process of individualization of mass-market goods and services which are used to fulfill particular customer needs at an affordable and reasonable price. They discussed the lack of flexibility in conventional enterprise resource planning and supply chain management systems to cope with the new requirements of market. To overcome these drawbacks, multi-agent systems are used to provide suitable means and also presented a solution to handle customization and corresponding information logistics in flexible way and to extend inter-business relations by partial automatic management.

In [44], the authors discussed the cost reduction and complex optimization problem for dispatching and planning scheduling of freight in a highly competitive market with an increasing share for general cargo. To minimize the complexity in such scenario, an autonomous coordination of transport services and planning processes can help. The paper introduced a multi-agent based approach that solved mentioned issues by enabling an autonomous dispatch process in mentioned scenario and also tackled resource allocation problems. Additionally, it supports a human dispatch manager in decision by developing a Decision-Support System (DSS). The responsibility of DSS is to provide proposals for allocations of transport orders to trucks. Table I discusses the role of multi-agent systems for the sub-functions of logistics with the proposed solution. Different multi-agent systems are proposed to solve different issues like optimization in the food supply chain, automation problem for loading-unloading, e-commerce solution for customised services, resource allocation problem, mass-customization, information management issue and to reduce cost and complex optimization.

TABLE I: Multi-Agent Systems in Logistics

Article	Targeted Problem	Solution
[29]	Optimization problem in food supply chain	Agent technology model to optimize multiple tasks like bullwhip effect, inventor and communication issues
[21]	Automation problem for loading and unloading	Automatic system to reduce cost of labor and consumption time
[36]	E-commerce for supply chain	Multi-agent system for an e-commerce environment to keep the stability of its own business process while providing customized services to the supply chain members
[37], [38], [40]–[42]	Resource allocation problem	Multi-agent system for traffic modeling, decision support system, logistic planning and transportation scheduling
[41]	Mass customization and information management issue	Co-operative multi-agent system to handle mass customization and extend inter-business relationship
[44], [45]	Cost reduction and complex optimization problem	Multi-agent mechanism of autonomous coordination of transport services for dispatching and planning scheduling

B. Robotics

In order to optimize internal material flow, the requirement and demand of industrial robot-technologies is increasing. The use of robotics and other technologies in logistics is common and research institutes (like Massachusetts Institute of Technology, the Bremen Institute of Production and Logistics (BIBA), the Institute of Shipping Economics and Logistics (ISL), London Business School, etc.) are trying to integrate different technologies in logistics. In [20], authors classified robotics-logistics activities in many scenarios such as loading/unloading and palletizing/depalletizing of goods, and discussed possible scenarios for research and development activities. Companies are trying to enhance functionality and flexibility of working in production units by means of robotics.

In [46] the authors discussed introduction of robots in industry. The idea was initially to use it for production of components of industrialized building and modular housing. In earlier Seventeenth century, first robots had been designed for construction and at the end the century, construction site had been developed. In Japan, it had been used to improve quality in prefabrication of modular homes. In addition to it, maintenance and safety robots had been developed for cleaning, inspection and safety. Furthermore, humanoid robots for construction are already tested, but service robots are in planning to build environment in future.

Nowadays, mobile robotics are significantly used not only for commercially and personally, but also for education and research due to offered new application [47]. An autonomous mobile mechatronic system for learning and research has been introduced by company Festo Didactic. This system is named as Robotino [48] and also introduced as standardized platform for education in the area of engineering and information technology. In [49], the authors discussed the impact of Robotino. They argued that this system provides an easy methods for education and also creates enabling environment to the industrially-relevant engineers for practical training.

Autonomous robots contain a software component and perform task-level executions depending upon the instructions for a certain goal. Planning in this case is still an exception rather than the norm because domains are often too dynamic or complex. In [50], authors characterized the RoboCup Logistics League (RCLL) as a medium complex robotics planning domain considering properties, implementation strategies, and planning models and also proposed a RCLL testbed as a benchmark for comparison.

In e-commerce, logistics requires attention to handle problems such as delays in deliveries or wrong deliveries, packages lost while shipping or damaging goods due to improper packing or handling. Automation helps to improve efficiency of storage and retrieval system. Furthermore, it extends their capacities and capabilities through autonomous storage and retrieval system (ASRS. But it has limitation of flexibility due to variant properties of order. In order to get a balance between efficiency, scalability and flexibility, the use of robotics is vital. In [51], the authors discussed the use of mobile robots in industrial developments. These are in being used to achieve robotic picking methods and extended the Product Service System (PSS) where the prime goal is to focus core competencies using a Logistics Automation Service System (LASS) business model.

Table II discusses the problems of loading/unloading, palletizing/depalletizing, construction, education and research methods and training, planning system defects and e-commerce logistics bottle-neck and also lists the possible solution for the mentioned problems using robotics to get economical benefits by reducing labour cost, time and improving efficiency.

C. Online Technologies

In the era of Internet, customers migrated into the online world, therefore Internet marketers use emails to collect and organize data for potential prospects. Email becomes a primary way among many business marketers to connect with customers. With the emerging of e-commerce on the Internet, a new form of marketing has evolved. Online marketers use different strategies to get attention of their customers from online banners to pop ups. Classically, technical marketing focused on the design of product with the specifications and key features, but at the same time designed to appeal to customers with basic information of product. However, it has also grown marketing strategies to encompass any use of modern technology as a marketing tool. Considering a case of a software company, which offers Adobe Systems having worth billions of dollars. A wide range of products are offered by this system. A number of companies rely on it and ts potential customers are basically skilled, highly computer literate professionals. The use of this product is a marketing tool for marketers.

The paper [52] addressed the effective use of Information Technology (IT) capabilities and the synchronous online technology in education section. In a teaching, learning and developing environment of an institute, it presented a theoretical model considering key capabilities of IT and synchronous online technology to support the large scale deployment, e.g., Black-board Collaborate. Apart from the education sector,

TABLE II: Robotics in Logistics

Article	Targeted sub-task	Solution
[20]	Loading/unloading and palletizing/depalletizing of goods	Automation using robotics for economical efficiency and flexibility
[46]	Construction	Construction robots to increase quality in prefabrication of home, maintenance and safety robots had been developed for cleaning, inspection and safety
[47] [49]	Education and research	Standardized platform for education in engineering and information technology
[50]	Planning Systems for logistics	Proposed a RCLL testbed as a benchmark for comparison of planning systems
[51]	E-commerce logistics issues	Industrial developments, PSS and LASS business model for efficiency, scalability and flexibility

TABLE III: Online Technologies in Logistics

Article	Targeted Sub-area	Solution
[52]	Teaching and Learning	Theoretical model on basis of key IT capabilities and synchronous online technology to support for deployment on large scale
[53] [54]	Online-shopping behaviours	Analysis of shopping trend and identification of customer need from online data collect
[55]	Advertisement and Marketing	Technical marketing using enabling online/computer technologies
[56] [57]	E-commerce	Online shopping

these online technologies are also used to observe the general online shopping behavior and particular on gender or cultural base behaviour [53] [54]. In competitive business environment where we have a number of providers, the key point in the hand of organizations is to understand the desires and needs of their customers. These online technologies help to analyze the trend of customer behaviour and adopt new policies for their product according to customer need and desire.

D. Wireless Communication Technologies

The article [22] showed the importance of construction supply chain management (CSCM) and how its is helpful. Construction projects are complex and big in size, e.g., high rised buildings or mega-sized buildings. Usually, on the construction site, there is little storage space, but high demand of construction components and materials. Hence, for the success, efficient and optimized supply chain management is required. In-spite of the availability of research and development of radio frequency identification (RFID), the mobile devices are still required to carry to check logistic flow in supply chain process. However, by the use of RFID and Wireless Sensor Networks-based operations, the equipment can become main drive for the whole process and may include movers, trailers, gates, and hoists. Another article [23] discussed the use of Wireless Mesh Network (WMN) for logistics and Wireless Sensor Network (WSN) to control and manage the logistic flow. The enlisted activities are the logistic functions of handling, packaging and distribution.

E. Internet of Things (IoT)

Internet of Things (IoT) is enabling technology and researchers are focusing on enabling the material procurement process improvement of a manufacturer by using it. IoT and Cloud Manufacturing (CM) are linked technologies, are practically inadequate, particularly for a highly service-driven manufacturing execution system. In this system CM supports to respond in capturing the IoT-enabled execution hierarchy dynamically [58]. In the supply chain network, the One Stop Logistic Service Provider (1SLP) is an integrator, which is used to design and implement comprehensive solutions for logistic service by assembling the resources, capabilities, and integrated technologies of supply chain networks. In [59], authors developed the IoT enabled 1SLP process to reduce excessive operation times and integrate with the network information. The authors in [60] proposed a novel multilayered vehicular data cloud platform. This platform is designed by using cloud computing and IoT technologies and used for warranty analysis of vehicles in the IoT environment. IoT can provide a good backbone support for ITS, however this area is not mature yet.

The IoT could also contribute significantly in the food and agribusiness industry. In food supply, perishable products have unpredictable supply variations and mean time they require food safety and sustainability. In this scenario IoT are used to solve these issue because it allows for remotely controlling the location, conditions of shipments and products. In [61], authors also developed a reference architecture using IoT for logistic information systems where it supported the provision of affordable tailor-made solutions.

Apart from the production, handling and safety of machinery and material is also important. For material handling, authors in [62] explained material handling system via analysis and performance availability in regard to energy-harvesting, ultra-low-power devices. They also discussed details of the hardware platform including architecture and testbed. They paid particular attention to the inBin smart device and energy-harvesting in the mentioned system. For the safety, the authors in [63] developed a monitoring system, particularly focusing for the type of cranes used to hoist heavy loads in the open air environment with the help of IoT. The latter included both hardware unit and software, and applied in engineering.

F. VANET

In case of disaster areas, an important task is to manage the resources via restoring the information flow to help in the recovery process. VANET is a wireless technology that is deployed in such disaster areas to recover the communication link [64] [65]. Communication among vehicles is discussed in [25] and authors presented the concept of car-2-car communication as 'smart object' and aimed to increase driving with comfort and safety. This communication should be secure and

TABLE IV: Communication Technologies in Logistics

Article	Targeted Sub-task	Solution
[58]	Synchronization issue	A real-time production logistics synchronization system
[59]	Operation time and information flow	IoT enabled 1SLP process to reduce excessive operation times and integrate with the network information
[61]	Food and agribusiness industry	Reference architecture for IoT-based logistic information systems where it supported the provision of affordable tailor-made solutions
[62]	Material handling	Material handling systems in regard to energy-harvesting, ultra-low-power devices
[63]	Safety and monitoring	Monitoring system for heavy loads in the open air environment with safety monitoring system

confidential [66]. The article [24] discussed its support for applications that notify about route hazards and incidents to both the drivers and logistics coordinators. Apart from the ITS usage, VANET applications includes a number of scenarios, such as information dissemination for safety like emergency alerts, traffic condition, service messages, road jam due to accidents, and collision avoidance. The disseminated information can be used for fleet coordination or rerouting of vehicles [67], but also for advertisements (e.g. disseminate marketing data), multimedia content distribution (e.g. audio, video streaming) [68], [69], [70] and environmental monitoring networks to gather information data (e.g. pollution, traffic monitoring and road pavement defects) [71]. For future services, the expansion of the smart grid represents a unique challenge in terms of the convergence of network platforms. WiFi as VANETs have the potential to become a reliable wireless network platform to support both the requirements of the smart grid and ITS-based services. Logistics companies are required to have flexible and cost efficient system to monitor, control and deliver products on time. The customers are also interested to have hassle free dealing with the logistics companies. In all these cases, multi-hop wireless broadcast is an important component in vehicular networks because network properties allows to exploit this feature. VANET is a good choice to fit in these scenarios due to its unique characteristics, low cost and simple deployment.

In port scenarios,complexity is increasing with the expansion of supply chains. Modern ports require advanced track and trace, security, information sharing and monitoring and VANET has the potential to fulfill these needs of port facilities [72] [73]. In [74], authors discussed the adoption of communication technologies to experience high levels of visibility, control and connectivity across the entire supply chain and examined the feasibility of VANET in a multi-modal logistics environment. They recommended architecture to provide mentioned goals, which also assure security while accessing. An other article [75] explained its potential to manage the flow of goods and resources efficiently, particularly within international ports. In [76] authors discussed the key role of Information and Communication Technology (ICT in managing logistics operations and supply chains. Table V summarizes this discussion along target sub-tasks and solutions provided by the mentioned technology.

III. VANET Projects, Applications and Discussion

Under the European Commission, a wide variety of projects for transportation is currently underway. Intelligent Transport System (ITS) [77] aims to develop road safety and traffic management applications. Secure vehicular communication, passenger comfort and infotainment are also objectives of ITS. To generate novel ideas and development of novel technology, the project "Transport Research and Innovation Portal (TRIP)" [78] considers of transportation aiming to give an overview of research activities at European Union and National level. The European Commission also started a project in for "Transport Research and Innovation in Horizon 2020" [79] to generate ideas for growth of transportation, transport sustainability, seamless mobility and also viewing European Union (EU) as a leader on the globe. Car-2-Car is the project of Car-2-Car Communication Consortium (C2C-CC) [80]. This project aims to contribute to the reduction of deaths in road accidents, reduce traffic congestion, improve efficiency and reduce the impact of the traffic on the environment.

Many projects are also running to develop prototypes of VANET for industry. Car-to-car cooperation [81] is a VANET project running in the Aqua-lab of Northwestern University. This project aims to provide information and entertainment to the passengers and automotive safety, and to reduce the impact of traffic on environment and smooth traffic flow. The project "Innovative Wireless Technologies for Industrial Automation (HiFlecs)" [82] in the University of Bremen, develops innovative wireless technologies for industrial real-time closed-loop applications. In the future industry, the wireless technologies allows to connect machinery and control units wirelessly. There are different challenges for future Industry 4.0 applications like low latency, highly reliability, deterministic, and secure communications. To meet these challenges, HiFlecs develop key technologies for an industrial wireless communication system with new functionality and features for real-time control applications. Intelligent System and Sensors [83] is the project of Auto21 for the development of control and monitoring of vehicle behavior, guidance, navigation, telematics, driving assistance and automation. Another funded project [84] of Auto 21 named "Vehicle Communications And Applications" at the Interlab of University of Sherbrook focuses on the development and testing of cost effective communication infrastructure, design of cooperative control strategies and their integration for vehicular communication applications."Canadian Association of Road Safety and Professional (CARSP)" [85] is dedicated to enhance road safety by developing safety applications. Last, U.S. department of transportation is dealing with safety application where focused applications are emergency electronic brake lights, blind spot brakes, forward collision warnings, etc. It outlines new ITS Strategic Plan 2015-2019 [86] and provides a framework around with ITS Joint Program Office for research, development, and adoption activities to achieve goals. This plan is built around two key points i.e. priorities-realizing connected vehicle implementation and advancing automation. Furthermore, this plan includes program categories regarding connected

TABLE V: Current Role of VANET in Logistics

Article	Targeted Sub-task	Solution
[72] [73]	Enhancing port facilities	Provide clustering solution to fulfill needs of port facilities like track and trace, security, information sharing and visibility
[74]	Visibility, control and connectivity across entire supply chain	Provide enhances visibility and connectivity
[75]	Monitoring and coordination of portside vehicular traffic	Reliable applications for monitoring and coordination, Efficient solution for information sharing
[76]	Logistic operations at intra and inter-organizational level	Build communication links between enter-prises and for many organizations around the world

TABLE VI: Active Projects for Transportation

Project	Purpose	Organization
ITS [77]	Road safety,traffic management, secure vehicular communication	European Commission
TRIP [78]	Research activities for transportation	European Commission
Horizon 2020 [79]	Ideas generation and growth sustainability of transport	European Commission
C2C [80]	Fatalities reduction during road accidents,Improve efficiency,Reduce impact on environment	C2C-CC
Car-to-Car cooperation [81]	Automotive safety, infotainment and entertainment for passengers, reduce traffic impact on environment, smooth traffic flow	Aqua-lab,Northwestern University
HiFlecs [82]	To develop innovative wireless technologies for industry	University of Bremen
Intelligent system and sensors [83]	Control and monitoring of vehicle behavior,vehicle guidance navigation and telematics, driving assistance and automation	Auto21
Vehicular communication and applications [84]	Development and testing of cost-effective communication infrastructure, design of cooperative control strategies	Interlab, University of Sherbrook funded by Auto21
CARSP [85]	Road safety	CARSP
ITS 2015-2019 Strategic Plan [86]	Connected Vehicles, automation, emerging capabilities, enterprise data, inter-operability, accelerating deployment	U.S. Department of Transportation

TABLE VII: Technologies Integration in Logistics and Transportation

Technologies	Method	Purpose	Properties
Multi-Agent System [29] [87]	Centralized and decentralized	Chain optimization, e-market place	Flexibility, simplicity, transaction and harmonization
Online Technologies [88]	Business collaboration	E-business	Availability, simplicity
Robotics [20] [21]	Experisim, realism, inductivism	Automation	Economical efficiency, engin performance and flexibility
Wireless Sensor Networks (WSN) [22] [23]	Environmental monitoring, localization, controlling	Detection, process control, monitoring	automation in information flow and control
IoT [58] [62] [63]	Multilayered, centralized and decentralized	Synchronization, information flow, material handling and safety and monitoring	Real time synchronization, efficiency, flexibility
VANET [24] [25]	Vehicles control communication and road safety	Controlled and efficient traffic management, automation	Low latency, high reliability, deterministic and secure communication, cost-effective and flexibility

vehicles, automation, emerging capabilities, enterprise data, interoperability and accelerating deployment. Table VI gives the summary of the all presented projects where road safety or automotive safety are key targets. However, some projects are focusing on the monitoring and some projects are working to improve the traffic efficiency with minimal impact on the environment and cost effective communication infrastructure. From the current projects objectives, we conclude that companies are looking for low cost, automotive safety and monitoring solutions to support logistics and general transport applications with high reliability. Table VII summarizes the technologies integrated in logistics and transportation. Each technology aims to solve a problem by using specialized methods, and upon successful implementation and deployment, each system exhibits certain properties. As mentioned in the table, Multi-agent System [29] [87] are used for chain optimization and to create e-market place through centralized and decentralized method. By applying this technique, the system becomes more simple, flexible and harmonized but less coordinated. Online technologies [88] are used to maintain the system information for customers, re-sellers, business partners and to help in the online collaboration with partners, exchange of documents for contracts with customers, suppliers and also negotiation of contacts have become more easy. They provide a centralized and fast information management system, and also introduce e-business opportunities. Therefore, the advanced systems become more simple, flexible and also increase the accessibility and availability of data. Robotics [20] [21] is another technology, which renders the advanced system to be more flexible, efficient and cost effective. Robotics are used for automation of the system using different methodologies like expericism, realism and inductivism. For monitoring the environment, and controlling the process, WSN [22] [23] is applied. This technology allows for automation in information flow and control through environment monitoring, localizing the system, and controlling methods. To improve transportation efficiency, and for controlled and efficient traffic management automation, VANET technology [24] [25] [89] is imposed through communication methods. By using these wireless communication technologies, the transportation can become

more secure, highly reliable and cost effective.

Currently, as mentioned in previous sections, VANET focuses on both safety [90] [91] and non-safety applications [92]. Non-safety applications are used to create commercial opportunities by increasing the number of equipped vehicles with on-board wireless devices. To make journey more pleasant for travelers, comfort and infotainment applications are being design and developed to provide information support and entertainment. Vehicular networks can also be employed to provide connectivity to catastrophe hit areas or remote rural communities lacking a conventional communication infrastructure to provide connectivity. Furthermore, Vehicular applications (for

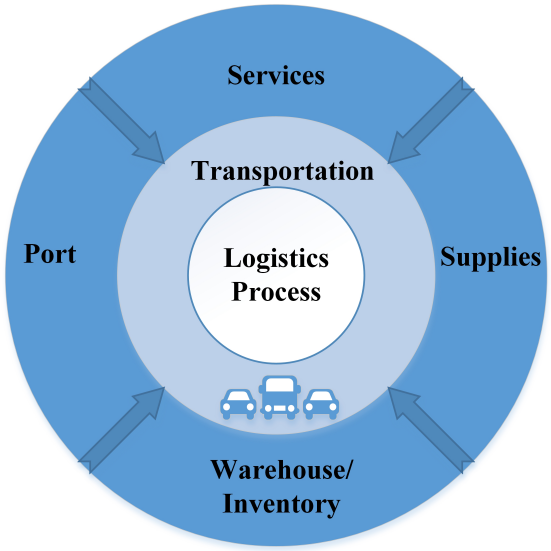


Fig. 4: VANET Integration in Logistics

example enhanced route guidance [93] and coordination by logistics providers, optimal scheduling of traffic light, and lane merging assistance by public coordinators) are intended to optimize routes [94], while also providing a reduction of gas emissions and fuel consumption. These applications are even more better when sensors are deployed for monitoring and controlling with VANET technology [95].

In logistics and transportation, many applications can be used like automatic vehicle detection or vehicle parking system for the logistic hot-spots (e.g sea port, warehouses etc.). It can also be used for the efficient traffic management for online delivery of logistics goods from production units to the distributed points or warehouses. Additionally, it allows automatic traffic control, re-routing the traffic in case of the traffic congestion [96] for improving just in time delivery and speed limit enforcement. Figure 4 shows the major four hot-spots of the logistics process where VANET technology can be used to maintain the information and logistics flow with the coordination of other technologies where transportation is the key factor that links industries to customers and consumers. We already know that for automation in manufacturing we use robotics and to maintain database of products and information online technologies and sensors are already deployed on different production units. By integrating these deployed technologies with VANET, we can improve efficiency in automation in terms of information flow and control. Considering Figure 4, within logistic processes the key goals are automation, monitoring, process control, chain optimization and information base, and currently multiple technologies are being used in each hotpots to achieve goals. In supplies, the need of facilities include monitoring, controlling, information and management, where sensors, WSN and online technologies are being used to fulfill their requirement. In warehouse and inventory management, there are different challenges like pelleting, packaging, storing, etc. In this scenario, handling, controlling, coordination and managing information are important tasks to complete, where robotics, WiFi technologies, sensors, database technologies or online technologies are being used for efficient handling. In ports, the challenges are to load and unload in optimal time, handling and information management. Here robotics, WSN, Multi-agent Systems and VANET are being used to cope with them. At the service level, companies are looking for solutions to ease the processes of purchasing, shipping, monitoring, tracking and controlling. They are using online technologies, multi-agent and sensors to deal with those processes. We observed in this case study, VANET has the potential for not only to inter-link these hotspots, but also to improve each unit. Hence, efficiency in delivery can improve economic factors for industries, and improve service level for customers and consumers while managing the freight transportation by providing best routing information, alerts for traffic jam or rerouting for congestion.

IV. Conclusions and Future Work

VANET is the one of the enabling technology in ITS that is used for road safety, traffic management and logistics applications. With the objective of the technologies and VANET integration in logistics and transportation, an overview of all technologies with logistics functions have been derived and summarized. Different technologies like multi-agent systems, online technologies, robotics, wireless sensor networks have been used to fast the process of manufacturing and to make the deliveries just in time over the last few years. We highlighted different challenges for each technology to provide flexible and effective solutions in terms of time and cost. We discussed VANET as key technology in logistics and transportation regarding challenges to cope with the mobility and short contact duration, where it is observed that VANET has potential to provide flexible and cost-effective solutions for logistics and transportation. It also has capability to make bridge to interlink hotspots of logistics process. For the deployment of this technology along attractive applications, a number of projects are running. This article also gave an overview of projects focusing respective extensions.

References

[1] J. Eckhardt and J. Rantala, "The Role of Intelligent Logistics Centres in a Multimodal and Cost-effective Transport System," *Procedia-Social and Behavioral Sciences, Elsevier*, vol. 48, pp. 612–621, 2012.

[2] E. Comission. (2015) Mobility and Transport. [Online]. Available: http://ec.europa.eu/transport/themes/index_en.htm

[3] D. Eckhoff, N. Sofra, and R. German, "A Performance Study of Cooperative Awareness in ETSI ITS G5 and IEEE WAVE," in *IEEE 10th Annual Conference on Wireless On-demand Network Systems and Services (WONS), 2013*, 2013, pp. 196–200.

[4] W.-F. Alliance, "Wi-Fi Protected Access: Strong, Standards-based, Interoperable Security for Todays Wi-Fi Networks," *White paper, University of Cape Town*, 2003.

[5] E. Dahlman, S. Parkvall, J. Skold, and P. Beming, *3G evolution: HSPA and LTE for mobile broadband*. Linacre House, Jordan Hill, Oxford, OX28DP: Second Edition, Academic press, 2010.

[6] S. Sesia, I. Toufik, and M. Baker, *LTE: the UMTS long term evolution*. The Atrium, Southern Gate, Chichester, West Sussex, PO19 8SQ, United Kingdom: Wiley Online Library, 2009.

[7] K. A. Khaliq, J. Pannek, and A. Qayyum, "Methodology for Development of Logistics Information and Safety System Using VANET," *Springer Lecture Notes on Logistics Proceedings of the 5th International Conference on Dynamics in Logistics LDIC*, pp. 185–195, 2016.

[8] M. Amadeo, C. Campolo, and A. Molinaro, "Enhancing IEEE 802.11p/WAVE to Provide Infotainment Applications in VANETs," *Ad Hoc Networks*, vol. 10, no. 2, pp. 253–269, 2012.

[9] R. G. Kasilingam, "Logistics and Transportation," *Great Britain: Kluwer Academic Publishers*, 1998.

[10] C. L. Weber and H. S. Matthews, "Food-miles and the Relative Climate Impacts of Food Choices in the United States," *Environmental science & technology*, vol. 42, no. 10, pp. 3508–3513, 2008.

[11] E. Ros, "Shelter Container Fit for Habitation with Extendible Inner Volume," Aug. 24 1993, uS Patent 5,237,784.

[12] S. Tayur, R. Ganeshan, and M. Magazine, *Quantitative Models for Supply Chain Management*. New York: Springer Science & Business Media, 2012, vol. 17.

[13] C. Gimenez, T. van der Vaart, and D. Pieter van Donk, "Supply Chain Integration and Performance: The Moderating Effect of Supply Complexity," *International Journal of Operations & Production Management*, vol. 32, no. 5, pp. 583–610, 2012.

[14] D. Prajogo and J. Olhager, "Supply Chain Integration and Performance: The Effects of Long-term Relationships, Information Technology and Sharing, and Logistics Integration," *International Journal of Production Economics, Elsevier*, vol. 135, no. 1, pp. 514–522, 2012.

[15] H. Stadtler, "Supply Chain Management: An Overview," in *Supply Chain Management and Advanced Planning, Springer*, 2015, pp. 3–28.

[16] A. P. C. Postelnicu and A. P. D.-C. Dabija, "Transfer and Diffusion of New Technologies Within the Supply Chain of Multinational Companies with Operations in RomaniaA Contemporary Approach," in *Geopolitics, Development, and National Security, Springer*, 2015, pp. 53–66.

[17] A. Musa, A. Gunasekaran, Y. Yusuf, and A. Abdelazim, "Embedded Devices for Supply Chain Applications: Towards Hardware Integration of Disparate Technologies," *Expert Systems with Applications, Elsevier*, vol. 41, no. 1, pp. 137–155, 2014.

[18] A. Musa, A. Gunasekaran, and Y. Yusuf, "Supply Chain Product Visibility: Methods, Systems and Impacts," *Expert Systems with Applications, Elsevier*, vol. 41, no. 1, pp. 176–194, 2014.

[19] S.-J. Chuu, "An Investment Evaluation of Supply Chain RFID Technologies: A Group Decision-making Model with Multiple Information Sources," *Knowledge-Based Systems, Elsevier*, vol. 66, pp. 210–220, 2014.

[20] W. Echelmeyer, A. Kirchheim, and E. Wellbrock, "Robotics-logistics: Challenges for automation of logistic processes," in *IEEE International Conference on Automation and Logistics (ICAL)*, 2008, pp. 2099–2103.

[21] W. Echelmeyer, A. Kirchheim, A. J. Lilienthal, H. Akbiyik, and M. Bonini, "Performance Indicators for Robotics Systems in Logistics Applications," in *IROS Workshop on Metrics and Methodologies for Autonomous Robot Teams in Logistics (MMARTLOG)*, 2011, p. 55.

[22] T.-H. Shin, S.-W. Yoon, and S. Chin, "A Construction Supply Chain Management Process with RFID/WSN-based Logistics Equipment," *Journal of Construction Engineering and Project Management*, vol. 2, no. 4, pp. 11–19, 2012.

[23] L. Evers, M. J. Bijl, M. Marin-Perianu, R. Marin-Perianu, and P. J. Havinga, "Wireless Sensor Networks and Beyond: A Case Study on Transport and Logistics," 2005.

[24] A. E. C. Mondragon and E. S. C. Mondragon, "Smart Grid and Wireless Vehicular Networks for Seaport Logistics Operations," in *19th ITS World Congress*, 2012.

[25] S. Eichler, C. Schroth, and J. Eberspächer, "Car-to-Car Communication," in *VDE-Kongress*. VDE VERLAG GmbH, 2006.

[26] D. Simchi-Levi, X. Chen, and J. Bramel, *The Logic of Logistics: Theory, Algorithms, and Applications for Logistics Management*, New York, Heidelberg, Dortrecht, London, 2013.

[27] A. Chiarini and A. Douglas, "The Impact of Logistics Solutions on Customer Satisfaction: An Exploratory Qualitative Study of Sanufacturing Companies," *Sinergie Italian Journal of Management*, pp. 255–270, 2015.

[28] W. W. Powell and K. Snellman, "The knowledge economy," *JSTOR journal of Annual Review of Sociology*, pp. 199–220, 2004.

[29] E. Mangina and I. P. Vlachos, "The Changing Role of Information Technology in Food and Beverage Logistics Management: Beverage Network Optimisation Using Intelligent Agent Technology," *Journal of Food Engineering, Elsevier*, vol. 70, no. 3, pp. 403–420, 2005.

[30] J. Stock, "International Journal of Physical Distribution & Logistics Management," *Marketing Intelligence & Planning*, vol. 10, no. 7, pp. 12–15, 1992.

[31] J. R. Stock, "Development and Implementation of Reverse Logistics Programs," in *Annual Conference Proceedings, Council of Logistics Management*, 1998.

[32] L.-H. Shih, "Reverse Logistics System Planning for Recycling Electrical Appliances and Computers in Taiwan," *Resources, Conservation and Recycling, Elsevier*, vol. 32, no. 1, pp. 55–72, 2001.

[33] E. Behmanesh and J. Pannek, "A Closed Loop Supply Chain Model with Flexible Delivery Path and Extended Random Path-based Direct Encoding," 2015.

[34] M. G. Avci and H. Selim, "A multi-agent system model for supply chains with lateral preventive transshipments: Application in a multinational automotive supply chain," *Elsevier Journal of Computers in Industry*, vol. 82, pp. 28–39, 2016.

[35] H. K. Chow, K. L. Choy, and W. Lee, "A dynamic logistics process knowledge-based system–An RFID multi-agent approach," *Knowledge-Based Systems, Elsevier*, vol. 20, no. 4, pp. 357–372, 2007.

[36] W. Ying and S. Dayong, "Multi-agent framework for third party logistics in E-commerce," *Journal of Expert Systems with Applications, Elsevier*, vol. 29, no. 2, pp. 431–436, 2005.

[37] D. Perugini, S. Wark, A. Zschorn, D. Lambert, L. Sterling, A. Pearce *et al.*, "Agents in Logistics Planning-experiences with the Coalition Agents Experiment Project," in *Proceedings of workshop at the Second International Joint Conference on Autonomous Agents and Multiagent Systems (AAMAS 2003). Melburne, Australia*, 2003.

[38] K. Zhu and A. Bos, "Agent-based design of intermodal freight transportation systems," in *NECTAR Conference*, 1999.

[39] B. Burmeister, A. Haddadi, and G. Matylis, "Application of multi-agent systems in traffic and transportation," *IEEE Proceedings-Software Engineering*, vol. 144, no. 1, pp. 51–60, 1997.

[40] J. L. Adler and V. J. Blue, "A cooperative multi-agent transportation management and route guidance system," *Transportation Research Part C: Emerging Technologies*, vol. 10, no. 5, pp. 433–454, 2002.

[41] I. J. Timm, P.-O. Woelk, P. Knirsch, H.-K. Tönshoff, and O. Herzog, "Flexible Mass Customisation: Managing Its Information Logistics Using Adaptive Cooperative Multi-agent Systems," in *Developments in Logistics and Supply Chain Management*. Springer, 2016, pp. 203–211.

[42] O. N. Granichin, P. Skobelev, A. Lada, I. Mayorov, and A. Tsarev, "Comparing Adaptive and Non-adaptive Models of Cargo Transportation in Multi-agent System for Real Time Truck Scheduling," in *IJCCI*, 2012, pp. 282–285.

[43] T. Sprodowski and J. Pannek, "Stability of distributed MPC in an intersection scenario," in *Journal of Physics: Conference Series*, vol. 659, no. 1. IOP Publishing, 2015, p. 012049.

[44] F. Arendt, O. Klein, and K. Barwig, "Intelligent Control of Freight Services on the Basis of Autonomous Multi-agent Transport Coordination," in *Springer:Logistics and Supply Chain Innovation*, 2016, pp. 313–324.

[45] N. Anand, R. van Duin, and L. Tavasszy, "Ontology-based Multi-agent System for Urban Freight Transportation," *International Journal of Urban Sciences*, vol. 18, no. 2, pp. 133–153, 2014.

[46] T. Bock, "Construction Robotics," *Springer Journal of Autonomous Robots*, vol. 22, no. 3, pp. 201–209, 2007.

[47] U. Karras, D. Pensky, and O. Rojas, "Mobile Robotics in Education and Research of Logistics," in *Workshop on Metrics and Methodologies for Autonomous Robot Teams in Logistics (IROS 2011)*, vol. 72, 2011.

[48] M. Bliesener, C. Weber, K. Kling, U. Karras, and D. Zitzmann, "Festo Robotino Manual," *Denkendorf: Festo Didactic GmbH & Co. KG*, 2007.

[49] H. Weinert and D. Pensky, "Mobile Robotics in Education and Student Engineering Competitions," in *IEEE AFRICON, 2011*, 2011, pp. 1–5.

[50] T. Niemueller, G. Lakemeyer, and A. Ferrein, "The Robocup Logistics League as a Benchmark for Planning In Robotics," in *WS on Planning and Robotics (PlanRob) at Int. Conf. on Aut. Planning and Scheduling (ICAPS)*, 2015.

[51] G. Q. Huang, M. Z. Chen, and J. Pan, "Robotics in Ecommerce Logistics," *HKIE Transactions*, vol. 22, no. 2, pp. 68–77, 2015.

[52] S. Low, J. Goh, S. K. Yeung, and I. Chia, "Building IT Capabilities to Deploy Large-Scale Synchronous Online Technology in Teaching and Learning," in *International Conference on HCI in Business, Government and Organizations, Springer*, 2016, pp. 531–544.

[53] R. Smith, G. Deitz, M. B. Royne, J. D. Hansen, M. Grünhagen, and C. Witte, "Cross-cultural Examination of Online Shopping Behavior: A Comparison of Norway, Germany, and the United States," *Journal of Business Research*, vol. 66, no. 3, pp. 328–335, 2013.

[54] M. Koufaris, "Applying the Technology Acceptance Model and Flow Theory to Online Consumer Behavior," *Information systems research*, vol. 13, no. 2, pp. 205–223, 2002.

[55] D. Bowie, A. Paraskevas, and A. Mariussen, "Technology-Driven Online Marketing Performance Measurement: Lessons from Affiliate Marketing," *International Journal of Online Marketing (IJOM)*, vol. 4, no. 4, pp. 1–16, 2014.

[56] A. R. Ashraf, N. Thongpapanl, and S. Auh, "The Application of the Technology Acceptance Model under Different Cultural Contexts: The Case of Online Shopping Adoption," *Journal of International Marketing*, vol. 22, no. 3, pp. 68–93, 2014.

[57] T. Escobar-Rodríguez and E. Carvajal-Trujillo, "Online Purchasing Tickets for Low Cost Carriers: An Application of the Unified Theory of Acceptance and Use of Technology (UTAUT) Model," *Tourism Management*, vol. 43, pp. 70–88, 2014.

[58] T. Qu, S. Lei, Z. Wang, D. Nie, X. Chen, and G. Q. Huang, "Iot-based real-time production logistics synchronization system under smart cloud manufacturing," *The International Journal of Advanced Manufacturing Technology*, vol. 84, no. 1-4, pp. 147–164, 2016.

[59] A. P. Hsu, W. Lee, A. J. Trappey, C. V. Trappey, and A.-C. Chang, "Using system dynamics analysis for performance evaluation of iot enabled one-stop logistic services," in *Systems, Man, and Cybernetics (SMC), 2015 IEEE International Conference on*. IEEE, 2015, pp. 1291–1296.

[60] W. He, G. Yan, and L. Da Xu, "Developing vehicular data cloud services in the iot environment," *IEEE Transactions on Industrial Informatics*, vol. 10, no. 2, pp. 1587–1595, 2014.

[61] C. Verdouw, R. Robbemond, T. Verwaart, J. Wolfert, and A. Beulens, "A reference architecture for iot-based logistic information systems in agri-food supply chains," *Enterprise Information Systems*, pp. 1–25, 2015.

[62] M. Roidl, J. Emmerich, A. Riesner, M. Masoudinejad, D. Kaulbars, C. Ide, C. Wietfeld, and M. Ten Hompel, "Performance availability evaluation of smart devices in materials handling systems," in *IEEE/CIC International Conference on Communications in China-Workshops (CIC/ICCC) (2014)*, 2014, pp. 6–10.

[63] H. D. Zhao, H. Z. Wang, G. N. Liu, C. Li, and M. H. Zhao, "The application of internet of things (iot) technology in the safety monitoring system for hoisting machines," in *Applied Mechanics and Materials*, vol. 209. Trans Tech Publ, 2012, pp. 2142–2145.

[64] Z. Alazawi, S. Altowaijri, R. Mehmood, and M. B. Abdljabar, "Intelligent Disaster Management System Based on Cloud-enabled Vehicular Networks," in *ITS Telecommunications (ITST), 2011 11th International Conference on*. IEEE, 2011, pp. 361–368.

[65] J. Sun, X. Zhu, C. Zhang, and Y. Fang, "RescueMe: Location-based Secure and Dependable VANETs for Disaster Rescue," *Selected Areas in Communications, IEEE Journal on*, vol. 29, no. 3, pp. 659–669, 2011.

[66] N. R. Siddiqui, K. A. Khaliq, and J. Pannek, "VANET Security Analysis on the Basis of Attacks in Authentication," *Springer Lecture Notes on Logistics Proceedings of the 5th International Conference on Dynamics in Logistics*, pp. 491–502, 2016.

[67] A. Rasheed, H. Zia, F. Hashmi, U. Hadi, W. Naim, and S. Ajmal, "Fleet & convoy management using VANET," *Journal of Computer Networks*, vol. 1, no. 1, pp. 1–9, 2013.

[68] M. S. Akbar, K. A. Khaliq, and A. Qayyum, "Vehicular MAC Protocol Data Unit (V-MPDU): IEEE 802.11p MAC Protocol Extension to Support Bandwidth Hungry Applications," in *Vehicular Ad-hoc Networks for Smart Cities, Springer*, 2015, pp. 31–39.

[69] K. A. Khaliq, J. Pannek, and A. Qayyum, "Suitability of IEEE 802.11ac/n/p for Bandwidth Hungry and Infotainment Applications for Cities," *IEEE SAI Intelligent Systems (IntelliSys 2016)*, pp. 499–509, 2016.

[70] M. S. Akbar, M. S. Khan, K. A. Khaliq, A. Qayyum, and M. Yousaf, "Evaluation of IEEE 802.11n for Multimedia Application in VANET," *Procedia Computer Science*, vol. 32, pp. 953–958, 2014.

[71] M. H. Arbabi and M. Weigle, "Using Vehicular Networks to Collect Common Traffic Data," in *Proceedings of the sixth ACM international workshop on VehiculAr InterNETworking, ACM*, 2009, pp. 117–118.

[72] A. E. C. Mondragon, E. S. C. Mondragon, and C. E. C. Mondragon, "Clustering DSRC-Based Networks for Logistics Operations in Ports," in *17th ITS World Congress*, 2010.

[73] C. S. Lalwani *et al.*, "Wireless Vehicular Networks to Support Road Haulage and Port Operations in a Multimodal Logistics Environment," in *IEEE/INFORMS International Conference on Service Operations, Logistics and Informatics (SOLI'09)*, 2009, pp. 62–67.

[74] A. E. C. Mondragon, C. S. Lalwani, E. S. C. Mondragon, and C. E. C. Mondragon, "Facilitating Multimodal Logistics and Enabling Information Systems Connectivity Through Wireless Vehicular Networks," *Elsevier International Journal of Production Economics*, vol. 122, no. 1, pp. 229–240, 2009.

[75] A. E. C. Mondragon, E. S. C. Mondragon, C. E. C. Mondragon, and F. Mungau, "Estimating the Performance of Intelligent Transport Systems Wireless Services for Multimodal Logistics Applications," *Elsevier Expert Systems with Applications*, vol. 39, no. 4, pp. 3939–3949, 2012.

[76] A. E. C. Mondragon, E. S. C. Mondragon, and C. E. C. Mondragon, "Innovative Information and Communication Technology for Logistics: The Case of Road Transportation Feeding Port Operations and Direct Short Range Communication Technology," in *Springer Supply Chain Management and Knowledge Management*, 2009, pp. 254–268.

[77] European Comission. (2015) Intelligent Transport System (ITS). [Online]. Available: http://ec.europa.eu/transport/themes/its/index_en.htm

[78] ——. (2015) Transport Research and Innovation Portal (TRIP). [Online]. Available: http://ec.europa.eu/transport/themes/research/trip_en.htm

[79] ——. (2015) Transport Research and Innovation in Horizon 2020. [Online]. Available: http://ec.europa.eu/transport/themes/research/horizon2020_en.htm

[80] (2015) Car-2-Car Communication Consortium (C2C-CC). [Online]. Available: https://www.car-2-car.org/index.php?id=5

[81] Northwestern University. (2015) Car-to-Car Cooperation. [Online]. Available: http://www.aqualab.cs.northwestern.edu/projects/111-c3-car-to-car-cooperation-for-vehicular-ad-hoc-networks

[82] Universität Bremen, Institut für Telekommunikation und Hochfrequenztechnik. (2015) Innovative wireless technologies for industrial automation (HiFlecs). [Online]. Available: http://www.ant.uni-bremen.de/en/projects/hiflecs/

[83] Auto21. (2015) Intelligent Systems and Sensors. [Online]. Available: https://auto21.ca/en/subcontent?page=ae2600

[84] Interlab, University of Sherbrook, Canada. (2015) Vehicle Communications And Applications. [Online]. Available: http://www.gel.usherbrooke.ca/interlab/index.php?page=projects

[85] (2015) Canadian Association of Road Safety Professionals (CARSP). [Online]. Available: http://www.carsp.ca/

[86] U. D. of Transportation. (2016) ITS Strategic Plan 215-16 . [Online]. Available: http://www.its.dot.gov/research_areas/strategicplan2015.htm

[87] M. Giannakis and M. Louis, "A Multi-agent Based Framework for Supply Chain Risk Management," *Journal of Purchasing and Supply Management, Elsevier*, vol. 17, no. 1, pp. 23–31, 2011.

[88] F. Lai, D. Li, Q. Wang, and X. Zhao, "The information technology capability of third-party logistics providers: a resource-based view and empirical evidence from China," *Journal of Supply Chain Management, Wiley Online Library*, vol. 44, no. 3, pp. 22–38, 2008.

[89] X. Yan, P. Yi, D. Zhu, and L. Fu, "ICTIS 2013: Improving Multimodal Transportation Systems-Information, Safety, and Integration." American Society of Civil Engineers, 2013.

[90] K. A. Hafeez, L. Zhao, Z. Liao, and B. N.-W. Ma, "Impact of mobility on vanets' safety applications," in *IEEE Global Telecommunications Conference (GLOBECOM 2010)*, 2010, pp. 1–5.

[91] M. S. Akbar, A. Qayyum, and K. A. Khaliq, "Information Delivery Improvement for Safety Applications in VANET by Minimizing Rayleigh and Rician Fading Effect," in *Vehicular Ad-hoc Networks for Smart Cities, Springer*, 2015, pp. 85–92.

[92] M. Ahyar and R. F. Sari, "Performance Evaluation of Multi-Channel Operation For Safety And Non-Safety Application On Vehicular Ad Hoc Network IEEE 1609.4," *International Journal of Simulation-Systems, Science and Technology*, vol. 14, no. 1, pp. 16–22, 2013.

[93] M. Khanjary and S. M. Hashemi, "Route Guidance Systems: Review and Classification," in *Proceedings of the 6th Euro American Conference on Telematics and Information Systems, ACM*, 2012, pp. 269–275.

[94] K. Collins and G.-M. Muntean, "A Vehicle Route Management Solution Enabled by Wireless Vehicular Networks," in *IEEE INFOCOM Workshops*, 2008, pp. 1–6.

[95] M. Ferreira and P. M. d'Orey, "On the Impact of Virtual Traffic Lights on Carbon Emissions Mitigation," *IEEE Transactions on Intelligent Transportation Systems*, vol. 13, no. 1, pp. 284–295, 2012.

[96] S. Dornbush and A. Joshi, "Street Smart Traffic: Discovering and Disseminating Automobile Congestion Using VANET's," in *IEEE 65th Vehicular Technology Conference (VTC2007-Spring)*, 2007, pp. 11–15.

WQbZS: Wavelet Quantization by Z-Scores for JPEG2000

Jesús Jaime Moreno-Escobar*, Oswaldo Morales-Matamoros*, Ricardo Tejeida-Padilla*,
Ana Lilia Coria-Páes*, and Teresa Ivonne Contreras-Troya†
*Instituto Politécnico Nacional, México.
†Universidad Autónoma del Estado de México, México.

***Abstract*—In this document we present a methodology to quantize wavelet coefficients for any wavelet-base entropy coder, we apply it in the particular case of JPEG2000. Any compression system have three main steps: Transformation in terms of frequency, Quantization and Entropy Coding. The only responsible for reducing or maintaining precision is the second element, Quantization, since it is the element of lossy compression that reduces the precision of dequantized pixels in order to make quantized pixels more compressible. We modify the well-known dead zone scalar Quantization introducing Z-Scores in the process. Thus, Z-scores are expressed in terms of standard deviations from their means. Resultantly, these z-scores have a distribution with a mean of 0 and a standard deviation of 1, in this way we increase redundancies into the image, which produces a lower compression ratio.**

Keywords*—*Z-Scores; Statistical Normalization; Wavelet Transformation; Scalar Quantization; Deadzone Quantization; JPEG2000

I. Introduction

One of the most important features of human beings is *Vision*, because is one of the most difficult sense to model, since not only involve mathematical models but also experience passages of the live of the a person, which can be different of each one. So, when a light ray enters into our eyes launches a highly complex process, which finalizes in the brain specifically into the visual cortex. Thus, scientists in this field intent to give a mathematical response of some of these features of the Human Visual System(HVS).

Digital image compression is a research topic for many years until today and a number of image compression algorithms is created for different applications. The JPEG2000 is a standard that tries to reduce the rate of stored pixels regarding its distortion rate, taking in account objective and subjective image quality. Several works have been demonstrated that the overall performance of JPEG2000 is superior to existing standards, as well as to supply functionality [1]. Figure 1 shows the comparison of bit rate of (a) JPEG and (b) JPEG2000 image coders, tested with the 24-bit and 512×512 pixel Color Image *Lena*. The results for this particular case and bit rate show that JPEG2000 is better in 2.63 dB, which is an important improvement.

However, JPEG2000 barely provide relevant features of the human visual system, since for removing pixels, in order to find more redundancies inside the image, JPEG2000 mainly applies criteria of the Information Theory such as thresholds, for instance. This this lack of information introduces artifacts into the recovered image, which are notorious at high compression rates, that is because many the most visible pixels regarding its perceptual significance have been eliminated.

In addition, JPEG2000 s an image compression standard approach, which was proposed by ISO/IEC. from previous standards, also, it was created as a framework where the image compression system can have the behavior of an image processing algorithm. The decision on several important compression features such as quality or resolution are created after the generation of the coded codestream. Thus, JPEG2000 decoded many image algorithms from a single coded file, give as a result different chances for coded domain processing. As in any compression or coding system, Quantization procedure is one of the critical steps of JPEG2000 image coding and decoding algorithm. Many of the desirable properties of the JPEG2000 standard contain manly two quantization methods, such as Embedded Scalar Quantization.

In this paper, we provide an overview of well-known methods in addition to propose a new one WQbZS. This paper is organized as follows. In Section 2, an introduction to quantization methods used in JPEG2000 is presented. Section 3 describes the WQbZS algorithm implemented in JPEG2000, in addition, we define the statistical relevancy of the Z-Scores, and Section 4 provides the experimental results.

II. A Brief Description of JPEG2000 Quantization and its Environment

A. Image Compression System

General Theory of Systems defines $information$ as $-entropy$, i.e. $negentropy$. Let us define first the concept of *entropy*, which is the tendency a system has when it tend to disintegrate by itself or by external factors[2]. Thus, entropy means the grade of disorder of a system. In the same way, a recovered image should have almost the same total entropy as the original one, but using less bits per every pixel. That is, a compressed image should have more entropy per bit than the original one. In addition, let us to mention that the main objective of recent image compression systems is to increase redundancies of images, understanding that some frequencies are redundant. These redundancies inside frequencies can be obtained by statistical procedures or the estimation of visual irrelevancies[3, Sec. 1.2].

A system is generally defined as a subset mainly composed by three subsystems: an input, a process, and an output, in

(a) JPEG: 0.22bpp, PSNR = 27.39 dB

(b) JPEG2000: 0.22bpp, PSNR = 30.03 dB

Figure 1: Comparison of bit rate of JPEG2000 and JPEG image coders, tested with the 24-bit Color Image *Lena*.

some cases we can include a fourth subsystem feedback, these subsystems define a cybernetic model, which is depicted in Figure 2. Hence, any system is defined as a set of the three o four elements standing in interrelation among themselves and also with the environment of the system.

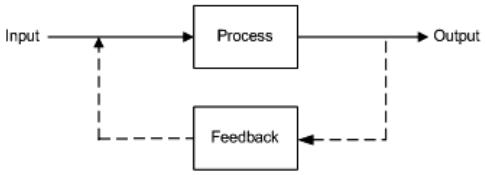


Figure 2: Description of *System* according to the General Theory of Systems.

Subsystems *Process* and *Feedback* have no relation, but *Feedback* is employed in order to adjust some characteristics or to evaluate how efficient is the *Process*. In the same way, an image compression coder is described as a general system as follows, Figure 3:

- *Input*: Original image considered with infinite and unquestionable quality $f(i,j)$;
- *Process*: Set of subsystems, these are: Forward Transformation, Quantization, Entropy Coding, Entropy Decoding, Inverse Quantization and Inverse Transformation;
- *Output*: Recovered image $\hat{f}(i,j)$;
- *Feedback*: Assessment of the possible distortion between original and recovered images, in order to measure the efficiency of the image compression system.

B. Dead-zone Uniform Scalar Quantizer in JPEG2000

Marcellin et.al. give us a general overview in [4] of the uniform scalar quantizer. This kind of quantization process is described as a mathematical model that maps every pixel or coefficient into a particular energy, which maintain the entropy but reduces the compression ratio. This way, this quantization values are uniformly distributed by rage known as a *QStep* with size Δ, this is fulfilled across the range, except when the energy of the pixel is quantized to zero, which is known as Dead-Zone. The width of this Zone extends from $-\Delta$ to $+\Delta$. Thus, a dead-zone can be defined as the quantization range around 0, which is twice the size of Δ, namely all the coefficients or pixels lower than $|\Delta|$ cannot be recovered in the dequantization process.

Thus, in a given wavelet plane ω_s^o , with spatial frequency s and spacial orientation o, and a particular *QStep* size Δ_s^o is used to quantize all the coefficients inside a wavelet decomposition. Hence a particular *QStep* is defined as follows:

$$\overline{c_{i,j}} = sign(c_{i,j}) \left\lfloor \frac{|c_{i,j}|}{\Delta_s^o} \right\rfloor \tag{1}$$

where $c_{i,j}$ is the original wavelet coefficient value, $sign(c_{i,j})$ denotes the sign of $c_{i,j}$ and $\overline{c_{i,j}}$ is the resulting *QStep* coefficient. Figure 4 illustrates such a *QStep* size Δ, here vertical lines indicate the thresholds of the quantization ranges and heavy points show the recovered coefficient.

The inverse quantizer or the recovered $\widehat{c_{i,j}}$ is given by

$$\widehat{c_{i,j}} = \begin{cases} (c_{i,j} + \delta)\Delta_s^o, & c_{i,j} > 0 \\ (c_{i,j} - \delta)\Delta_s^o & c_{i,j} < 0 \\ 0, & c_{i,j} = 0 \end{cases} \tag{2}$$

where δ is a parameter that intents to reconstruct $c_{i,j}$ at the center of a given quantization interval and varies form 0 to 1.

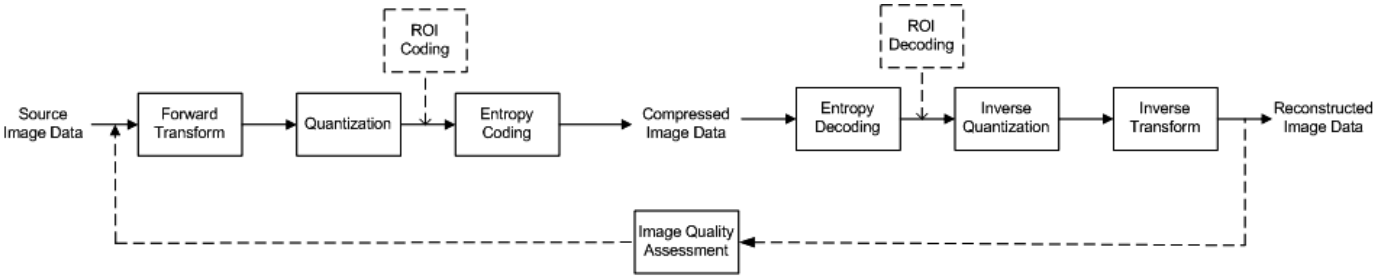


Figure 3: Scheme in order to define any Image Compression System.

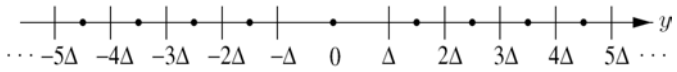


Figure 4: Dead-zone uniform scalar quantizer with *QStep* size Δ.

The International Organization for Standardization recommends to adopt the middle point in order to reconstruct $c_{i,j}$, setting $\delta = 0.5$ [1]. Pearlman and Said in [5] find that $\delta = 0.375$ obtains better results, especially for high frequency wavelet planes. when $-\Delta < y < \Delta$, the quantizer level and reconstruction value are both 0. Since it is known that many coefficients in a wavelet transform are close to zero (usually those of higher frequencies), it means that they can be on the dead-zone, namely $\widehat{c_{i,j}} = 0$.

It is important to realize that Quantization Process is the only subsystem that induces degradations into the image compression system, so when a wavelet plane ω_s^o is quantized is because $f(i,j)$ would be losslessly compressed, but the induction of $|\Delta| \neq 1$ causes a lossy compression, on the contrary when $|\Delta| = 1$ it causes lossless compression.

III. WQbZS Algorithm

A. Methodology

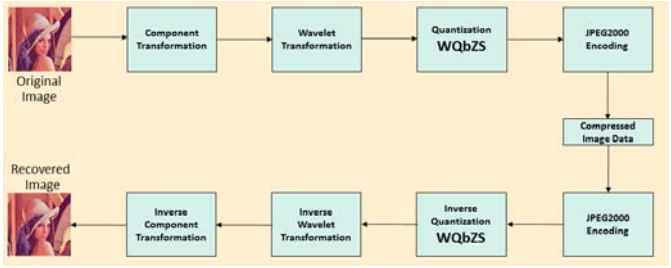


Figure 5: WQbZS Algorithm.

Figure 5 shows the present proposal, which is divided in 2 main stages with 3 steps each one:

1) Coding a JPEG2000 Image.
   - Component Transformation,
   - Wavelet Transform, and
   - Definition of Z-Scores for JPEG2000 Quantization.
2) Decoding a JPEG2000 Image.
   - Inverse Component Transformation,
   - Inverse Wavelet Transform, and
   - Definition of Z-Scores for JPEG2000 Inverse Quantization.

B. Component Transformation

For widespread applications , Data Compression systems of Natural Images, like JPEG2000, usually code color images. These images are numerically represented in several Chromatic Spaces both receptional representations, such as RGB or $CMYK$, and post-receptional representations, such as YC_bC_r, YCM, or HSB, being RGB the most commonly used along with YC_bC_r.

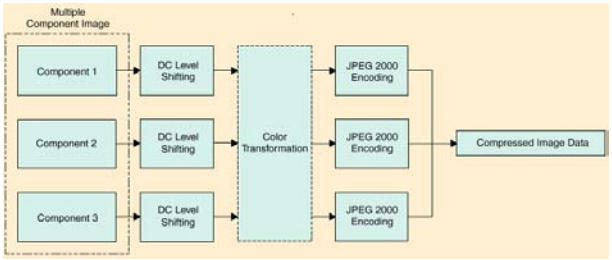


Figure 6: JPEG2000 Multiple Component Encoder.

In this way, a color image represented by a RGB color space, is decomposed into the same number of wavelength that in our cones in our retina can perceive, namely Red, Green, and Blue color components or cones. Figure 6 shows a special implementation of the post-receptional color space YC_bC_r, when JPEG2000 performs a chromatic coding, a complete encoding is performed at each color component. R, G and B color channels are numerically more dependent than Y, C_r and C_b, thus the chrominance channels are independently coded at lower size than luminance one in order to get better compression rates [6].

The JPEG2000 standard considers both Reversible Component Transformation (RCT) and Irreversible Component Transformation(ICT) [1, Annex G]. For lossy coding or with some degradations is employed an ICT, which makes use of the the 9/7 wavelet transform also irreversible. Thus, the forward and inverse filters to compute a 9/7 wavelet transform are estimated by the Equation 3 and 4, respectively [7], [3].

$$\begin{bmatrix} Y \\ C_b \\ C_r \end{bmatrix} = \begin{bmatrix} 0.299 & 0.587 & 0.114 \\ -0.16875 & -0.33126 & 0.5 \\ 0.5 & -0.41869 & -0.08131 \end{bmatrix} \begin{bmatrix} R \\ G \\ B \end{bmatrix} \quad (3)$$

$$\begin{bmatrix} R \\ G \\ B \end{bmatrix} = \begin{bmatrix} 1.0 & 0 & 0.114 \\ 1.0 & -0.34413 & -0.71414 \\ 1.0 & 1.772 & 0 \end{bmatrix} \begin{bmatrix} Y \\ C_b \\ C_r \end{bmatrix}. \quad (4)$$

RCT is commonly employed not only for lossy compression but also for lossless encoding, along with the 5/3 wavelet transform, which is also reversible. The forward RCT transformation is computed by Equation 5 while the inverse by the Equation 6.

$$\begin{bmatrix} Y \\ C_b \\ C_r \end{bmatrix} = \begin{bmatrix} \left\lfloor \frac{R+2G+B}{4} \right\rfloor \\ R-G \\ B-G \end{bmatrix} \begin{bmatrix} R \\ G \\ B \end{bmatrix} \quad (5)$$

$$\begin{bmatrix} R \\ G \\ B \end{bmatrix} = \begin{bmatrix} Y - \left\lfloor \frac{C_r+C_b}{4} \right\rfloor \\ C_b+G \\ C_r+G \end{bmatrix} \begin{bmatrix} Y \\ C_b \\ C_r \end{bmatrix} \quad (6)$$

IV. Wavelet Transform

The original image $\mathcal{I}$ used by JPEG2000 is separated into different spatial frequencies and orientations using a multi level or multiresolution Direct Wavelet Transform (DWT) either Reversible or Irreversible [8], [9], by each channel. Thus $\mathcal{I}$ is separated in different set of planes or wavelet planes ω or spatial frequencies, where each wavelet plane has details at a given spatial resolutions and it is defined as follows:

$$DWT\left\{\mathcal{I}\left(w\right)\right\} = \sum_{s=1}^{n} \sum_{o=v,h,d} \omega_s^o + c_n\left(t\right) \quad (7)$$

where $s = 1, \ldots, n$, n the number of wavelet planes (in frequency domain) and $c_n\left(t\right)$ the residual plane, it is important to mention that this plane is the only part of the DWT which is the time or pixel domain. Spatial orientation is represented by $o = v, h, d$ i.e. vertical, horizontal or diagonal details, respectively.

A DWT filters each row and column of $\mathcal{I}\left(w\right)$ with a high-pass and low-pass filters, respectively. This algorithm yields in the duplication of samples, so the resultant image is downsampled by 2 both for column and rows, thus the number of sample remains in the same amount.

The Reversible Direct Wavelet Transform is performed by a 5/3 filter. Analysis and its respective synthesis filters are exposed in Table I. While, 9/7 filter is employ to perform a Irreversible Direct Wavelet Transform and its analysis and synthesis filters are depicted by Table II.

It is invariant if the columns or the rows of the Y, C_r or C_b channels are processed first.

The number of filtering levels or stages n of wavelet planes, depends directly on its usage. Notwithstanding that, some authors report that the best results are gotten with $n = 3$ [10], if it is taking into account the trade-off between image quality and compression ratio,.

Figure 7 depicts the Irreversible Direct Wavelet Transform of the Y component applied in the image *Peppers* with $n = 3$.

Table I: 5/3 Analysis and Synthesis Filter.

Analysis Filter		
i	Low-Pass Filter $h_L(i)$	High-Pass Filter $h_H(i)$
0	6/8	1
±1	2/8	-1/2
±2	-1/8	
Synthesis Filter		
i	Low-Pass Filter $h_L(i)$	High-Pass Filter $h_H(i)$
0	1	6/8
±1	1/2	-2/8
±2		-1/8

Table II: 9/7 Analysis and Synthesis Filter.

Analysis Filter		
i	Low-Pass Filter $h_L(i)$	High-Pass Filter $h_H(i)$
0	0.6029490182363579	1.115087052456994
±1	0.2668641184428723	-0.5912717631142470
±2	-0.07822326652898785	-0.05754352622849957
±3	-0.01686411844287495	0.09127176311424948
±4	0.02674875741080976	
Synthesis Filter		
i	Low-Pass Filter $h_L(i)$	High-Pass Filter $h_H(i)$
0	1.115087052456994	0.6029490182363579
±1	0.5912717631142470	-0.2668641184428723
±2	-0.05754352622849957	-0.07822326652898785
±3	-0.09127176311424948	0.01686411844287495
±4		0.02674875741080976

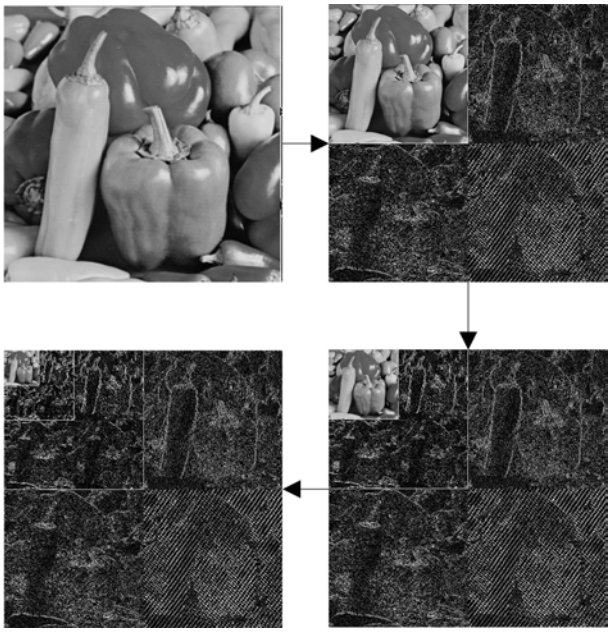


Figure 7: Three stages to decompose an image by means of a Direct Wavelet Transformation for *Peppers* image.

A. Definition of Z-Scores for JPEG2000 Quantization

A Z-score is a numerical measurement of a relationship of samples to the mean in a group of values. If a Z-score is 0, it represents the score is identical to the mean score. Exist another implementations of Z-scores and in some literature they are commonly known as the Altman Z-score. Edward Altman, a professor at New York University, developed and

introduced the Z-score formula in the late 1960s as a solution to the time-consuming and somewhat confusing process investors had to undergo to determine how close to bankruptcy a company was[11], for instance.

In this work we consider the samples as Intensity values either Y, C_r or C_b channel in wavelet domain. So, in this case Z-scores can be positives or negatives, with a positive value indicating the score is above the mean of the coefficients of wavelet planes and a negative score indicating it is below its mean. Positive and negative scores also reveal the number of standard deviations the score is either above or below the mean, namely if it is easy to increase the redundancies around the average of frequencies in different spacial orientations.

Z-scores also reveal if a wavelet decomposition is typical for a specified Image $\mathcal{I}(w)$ or if it is atypical. In addition to this, Z-scores also make it possible for analysts to adapt scores from various Images $\mathcal{I}(w)$ to make scores that are compared to one another accurately.

In this way, a z-score or standard score indicates how many standard deviations a coefficient ω_s^o is from the mean. The general Equation for estimate a z-score is calculated from Equation 8.

$$Z_s^o = \frac{\omega_s^o - \mu_s^o}{\sigma_s^o} \tag{8}$$

where Z is the Z-score, ω_s^o is the value of the coefficient in the wavelet domain, μ_s^o is the population mean, and σ_s^o is the standard deviation. s the number of a particular wavelet plane of ω^o in addition spatial orientation is represented by $o = v, h, d$ i.e. vertical, horizontal or diagonal details, respectively.

Z-scores of the the coefficients ω_s^o of a Direct Wavelet Transform can be interpreted as follows:

- A Z-score less than 0 represents a set of coefficients ω_s^o less than its mean μ_s^o.
- A Z-score greater than 0 represents a set of coefficients ω_s^o greater than its mean μ_s^o.
- A Z-score equal to 0 represents a set of coefficients ω_s^o to its μ_s^o.
- A Z-score equal to 1 represents a set of coefficients ω_s^o that is 1 standard deviation σ_s^o greater than its mean μ_s^o; a Z-score equal to 2, 2 standard deviations σ_s^o greater than its mean μ_s^o; etc.
- A Z-score equal to -1 represents a set of coefficients ω_s^o that is 1 standard deviation σ_s^o less than the mean μ_s^o; a z-score equal to -2, 2 standard deviations σ_s^o less than the meanμ_s^o; etc.

As the number of coefficient in the set ω_s^o is very large, about 68% of the set of coefficients ω_s^o have a Z-score between -1 and 1; about 95% have a Z-score between -2 and 2; and about 99% have a Z-score between -3 and 3.

In this way, when we introduce a Z-score expressed by Equation 8 to Equation 1, we propose Equation 9, which is a Z-score for JPEG2000 Quantization.

$$\overline{\omega_s^o} = sign(\omega_s^o) \left\lfloor \left| \frac{\omega_s^o - \mu_s^o}{\sigma_s^o} \right| \right\rfloor \tag{9}$$

Finally, we introduce $\overline{\omega_s^o}$ to a general decomposition expressed in Equation 7. Thus, we propose Equation 10 in order to quantized any wavelet coefficient subset ω_s^o, which will be encoded by JPEG2000.

$$\mathcal{I}(w) = \sum_{s=1}^{n} \sum_{o=v,h,d} sign(\omega_s^o) \left\lfloor \left| \frac{\omega_s^o - \mu_s^o}{\sigma_s^o} \right| \right\rfloor + c_n(t) \tag{10}$$

V. Experimental Results

We test our algorithm in two different ways. By one hand, we perform a test with a well-known image, *Lena* 8. By the other hand, we test our methodology with two important Image Databases, *CMU* and Image Databases*CSIQ*.

Nowadays, Mean Squared Error (MSE) is still the most used objective performance metrics and several quality assessments are based on it, Peak Signal-to-Noise Ratio (PSNR) is the best example of it. But some authors like Wang and Bovik in [12], [6] consider that MSE is a poor device to be used in quality assessment systems. In this work we use PSNR to compere our results regarding the ones obtained by the standard.

In this way, $f(i,j)$ and $\hat{f}(i,j)$ represent two images, which we want to compare and the size of them is the same. Then, $f(i,j)$ is the original reference image considered with unquestionable and perfect quality, in addition $\hat{f}(i,j)$ is a distorted version of $f(i,j)$. Then, the MSE and the PSNR are, respectively, defined as:

$$MSE = \frac{1}{NM} \sum_{i=1}^{N} \sum_{j=1}^{M} \left[f(i,j) - \hat{f}(i,j) \right]^2 \tag{11}$$

and

$$PSNR = 10 \log_{10} \left(\frac{{\mathcal{I}_{max}}^2}{MSE} \right) \tag{12}$$

where $\mathcal{I}_{max}$ is the maximum possible intensity value in $f(i,j)$ ($M \times N$ size). Thus, for images of 8 bits per pixel (bpp) per single channel $\mathcal{I}_{max} = 2^8 - 1 = 255$. Thus, for 24 bpp color images the PSNR is defined in the same way that in Equation 12, while MSE is the mean of individual MSE among Red, Green, and Blue channels, so once again $\mathcal{I}_{max} = 2^8 - 1 = 255$.

An important goal of any image compression systems is to improve the correlation of the pixels, since the higher correlation at the quantization, the more efficient coding system.

We employ PSNR because is a image quality assessment extensively used in the image processing field, since this metric have favorable features, such as:

1) A convenient metrics for the purpose of optimization of image coders. For example in JPEG2000, PSNR is employed both in Optimal Rate Allocation [13], [3] and Region of interest [14], [3].
2) By definition PSNR is the difference signal between the two compared images regarding the peak or the

maximum intensity error , namely $(\mathcal{I}_{max})^2$, giving a clear meaning of the energy of overall error inside a given signal.

A. *Image* Lena

Figure 8: 512×512 Image *Lena*.

Figure 9 shows that there is an important difference between the curves obtained by JPEG2000 and JPEG2000+WQbZS. In the particular case of Image *Lena*, JPEG2000 get, on the average, 36.6397dB when the image is compressed from 0.1 to 10.5 bits per pixel. While when the proposed algorithm is introduced to JPEG2000, performance of the standardized image compression system increases its performance, on the average, 3.7284 dB, which means that modifying the quantization step of the compressor reduces the error by half approximately.

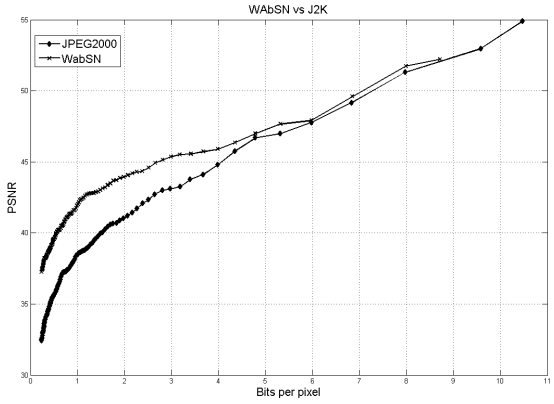


Figure 9: Bits per Pixel vs PSNR of Image *Lena*, Fig. 8.

B. CMU *Image Database*

This experiment is performed across the CMU Image Database (Annex A). Image quality estimations are assessed by PSNR.

Thus, the following experiments were performed on the selected images of *CMU* Image Database, which were transformed into YC_bC_r color representation, since it is the color space used by JPEG2000. Figure 10 shows the relation between compression rate and average quality. On average, any size image compressed by JPEG2000+WQbZS (dashed function) with 30 dB is stored at 0.65 bpp, while JPEG2000 (continuous function) stores it at 1.32 bpp.

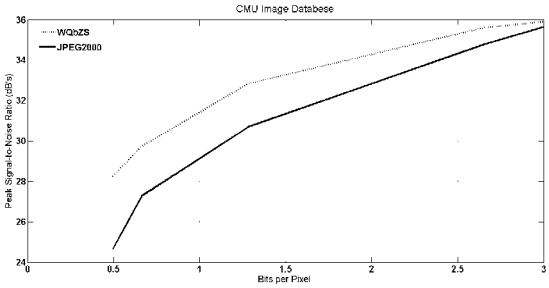


Figure 10: Comparison between JPEG2000 and JPEG2000+WQbZS image coders. Compression rate vs PSNR objective image quality, of the CMU Image database.

In Figure 11 we can see the difference when the image *Lena* is compressed at 0.5 bpp by JPEG2000 (a) and JPEG2000+WQbZS (b). At the same compression ratio, JPEG2000+WQbZS improves image quality by 1.04 dB. On average JPEG2000+WQbZS either compresses 0.27 bpp more with the same image quality or reduces in 0.93 dB the error with the same bit-rate.

C. CSIQ *Image Database*

This experiment is performed across the CSIQ Image Database (Annex B). Image quality estimations are assessed by PSNR.

Thus, the following tests are made on the selected images of *CSIQ* Image Database transformed into YC_bC_r color space (it is the color space used by JPEG2000). Figure 12 shows the relation between compression rate and average quality. On average, any size image compressed by JPEG2000+WQbZS (dashed function) with 35 dB is stored at 2.35 bpp, while JPEG2000 (continuous function) stores at 2.75 bpp.

In Figure 13 we can see the difference when the image *Bridge* is compressed at 0.35 bpp by JPEG2000 (a) and JPEG2000+WQbZS (b). At the same compression ratio, JPEG2000+WQbZS improves image quality by 1.15 dB. On average JPEG2000+WQbZS either compresses 0.28 bpp more with the same image quality or reduces in 0.947 dB the error with the same bit-rate.

VI. Conclusions

We defined Forward and Inverse statistical Quantizer using Z-score. We incorporated it to JPEG2000 proposing an alternative way dor the quantization step in the cited standard of image compression system. We expoused the mathematical explanation of normalizing or standardized the wavelet coefficients, since it increases redundancies, giving as a result a better image with the same entropy. In order to measure the effectiveness of the statistical quantization, a performance analysis is done using the image quality assessment PSNR, which measured the image quality between reconstructed and original images. Our results show that the employment of the Wavelet Quantization by means of Z-scores improves the JPEG2000 compression and image quality. In addition, when WQbZS is added into JPEG2000, it importantly improves the results getting the conventional JPEG2000 compression.

(a) JPEG2000: 0.5bpp, PSNR = 31.63 dB

(b) JPEG2000+WQbZS: 0.5bpp, PSNR = 32.67 dB

Figure 11: Comparison of bit rate of JPEG2000 and JPEG2000+WQbZS image coders, tested with the 24-bit Color Image *Lena*, taken from *CMU* Image Database.

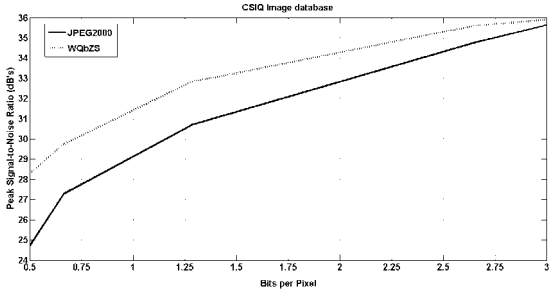


Figure 12: Comparison between JPEG2000 and JPEG2000+WQbZS image coders. Compression rate vs PSNR objective image quality, of the CSIQ Image database.

ACKNOWLEDGMENT

This work is supported by National Polytechnic Institute of Mexico (Instituto Politécnico Nacional, México) by means of Projects SIP-20160786, SIP-20161053, and SIP-20161713 the Academic Secretary and the Committee of Operation and Promotion of Academic Activities (COFAA) and National Council of Science and Technology of Mexico (CONACyT) by means of National Research System (Sistema Nacional de Investigadores) grants No. 56739 (Dr. Moreno), 32772 (Dr. Morales), and 335839 (Dr. Tejeida).

APPENDIX

A. University of Southern California Image Database

Figure 14 depicts the University of Southern California Image Data Base, *Miscellaneous volume*[15]. The database contains eight 256×256 pixel images and eight 512×512 pixel images [15].

B. Categorical Subjective Image Quality Image Database

CSIQ Database includes 30 original images (Figure 15), which are distorted by 6 different types of distortions at 4 or 5 degrees. CSIQ Database has 5000 perceptual evaluations of 25 observers[16].

REFERENCES

[1] M. Boliek, C. Christopoulos, and E. Majani, *Information Technology: JPEG2000 Image Coding System*, JPEG 2000 Part I final committee draft version 1.0 ed., ISO/IEC JTC1/SC29 WG1, JPEG 2000, April 2000.

[2] L. V. Bertalanffy, *Teoría General de los Sistemas*, M. Fondo de Cultura Ecocǿmica, Ed., 1989.

[3] D. S. Taubman and M. W. Marcellin, *JPEG2000: Image Compression Fundamentals, Standards and Practice*, ser. ISBN: 0-7923-7519-X. Kluwer Academic Publishers, 2002.

[4] M. W. Marcellin, M. A. Lepley, A. Bilgin, T. J. Flohr, T. T. Chinen, and J. H. Kasner, "An overview of quantizartion of JPEG2000," *Signal Processing: Image Communication*, vol. 17, no. 1, pp. 73–84, Jan. 2002.

[5] W. A. Pearlman and A. Said, "Image wavelet coding systems: Part II of set partition coding and image wavelet coding systems," *Foundations and Trends in Signal Processing*, vol. 2, no. 3, pp. 181–246, 2008.

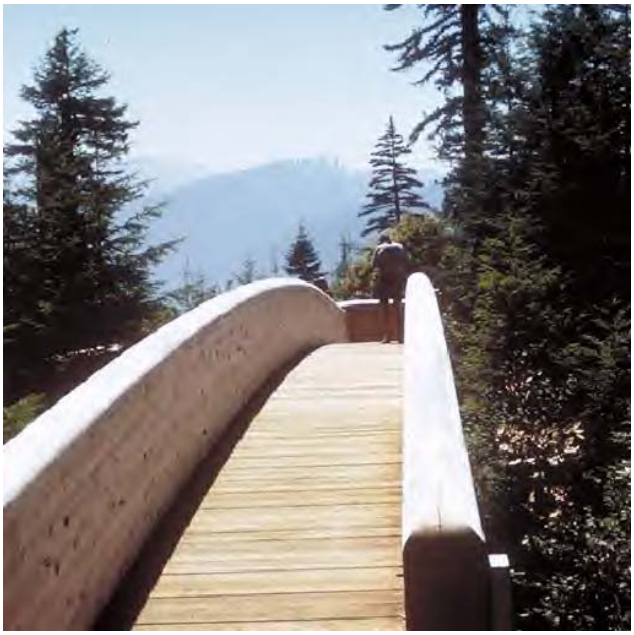
[6] Z. Wang and A. C. Bovik, *Modern Image Quality Assessment*, 1st ed. Morgan & Claypool Publishers: Synthesis Lectures on Image, Video, & Multimedia Processing, February 2006.

[7] A. Skodras, C. Christopoulos, and T. Ebrahimi, "The JPEG 2000 still image compression standard," *IEEE Signal Processing Magazine*, vol. 18, no. 5, pp. 36–58, September 2001.

[8] M. Antonini, M. Barlaud, P. Mathieu, and I. Daubechies, "Image coding using wavelet transform," *IEEE Transactions on Image Processing*, vol. 1, no. 2, pp. 205 – 220, April 1992.

[9] W. Sweldens, "The lifting scheme: A custom-design construction of biorthogonal wavelets," *Applied and Computational Harmonic Analysis*, vol. 3, no. 2, pp. 186 – 200, 1996.

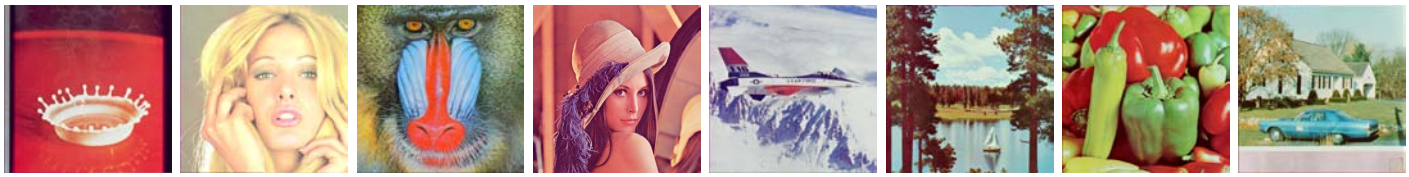
(a) JPEG2000: 0.35bpp, PSNR = 27.41 dB



(b) JPEG2000+WQbZS: 0.35bpp, PSNR = 28.56 dB

Figure 13: Comparison of bit rate of JPEG2000 and JPEG2000+WQbZS image coders, tested with the 24-bit Color Image *Bridge*, taken from *CSIQ* Image Database.

(a) Girl (b) Couple (c) Girl (d) Girl (e) House (f) Tree (g) Jelly beans (h) Jelly beans



(i) Splash (j) Tiffany (k) Baboon (l) Lena (m) F-16 (n) Sailboat (o) Peppers (p) House

Figure 14: Tested 24-bit Color Images, obtained from the University of Southern California Image Data Base. Figures (a) to (h) are 256×256 pixel Images, while Figures (i) to (p) are 512×512 pixel Images.

[10] A. Said and W. Pearlman, "A new, fast, and efficient image codec based on Set Partitioning In Hierarchical Trees," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 6, no. 3, pp. 243 – 250, June 1996.

[11] E. I. Altman, "Predicting financial distress of companies: Revisiting the z-score and zeta models," *Journal of Finance*, 1968.

[12] Z. Wang and A. Bovik, "Mean squared error: Love it or leave it? a new look at signal fidelity measures," *Signal Processing Magazine, IEEE*, vol. 26, no. 1, pp. 98 –117, jan. 2009.

[13] F. Auli-Llinas and J. Serra-Sagrista, "Low complexity JPEG2000 rate control through reverse subband scanning order and coding passes concatenation," *IEEE Signal Processing Letters*, vol. 14, no. 4, pp. 251 –254, april 2007.

[14] J. Bartrina-Rapesta, F. Auli-Llinas, J. Serra-Sagrista, and J. Monteagudo-Pereira, "JPEG2000 Arbitrary ROI coding through rate-distortion optimization techniques," in *Data Compression Conference*, 25-27 2008, pp. 292 –301.

[15] S. and Image Processing Institute of the University of Southern California. (1997) The USC-SIPI image database, available at http://sipi.usc.edu/database/. Signal and Image Processing Institute of the University of Southern California. [Online]. Available: http://sipi.usc.edu/database/

[16] E. C. Larson and D. M. Chandler, "Most apparent distortion: a dual strategy for full-reference image quality assessment," in *Proc. SPIE*, vol. 742, 2009.

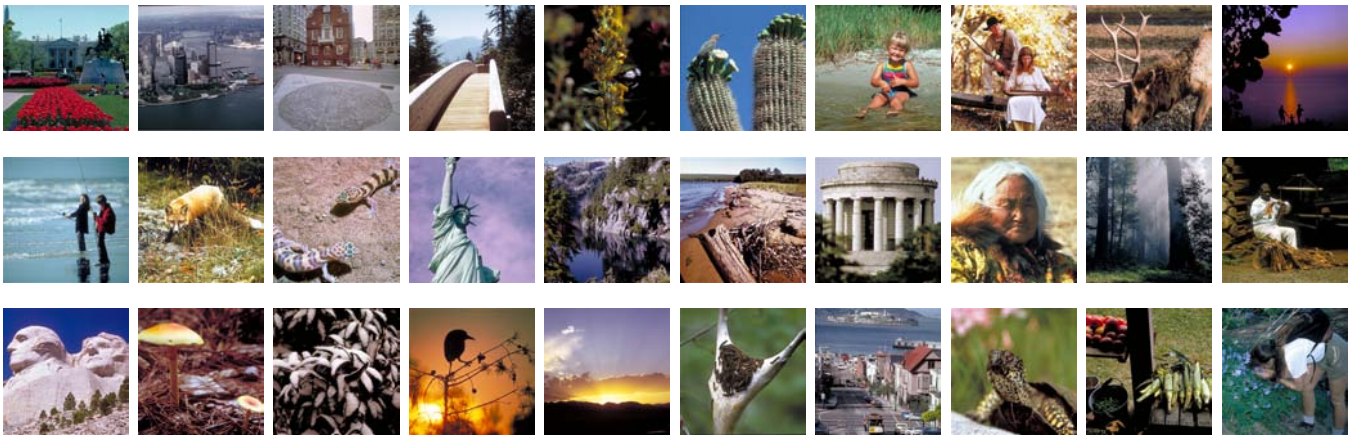


Figure 15: Tested 512×512 pixel 24-bit color images, belonging to the CSIQ Image database

Determination of Child Vulnerability Level from a Decision-Making System based on a Probabilistic Model

SAHA Kouassi Bernard
Laboratory of Computer Science and Telecommunications, National Polytechnic Institute
Abidjan, PO Box 475, Côte d'Ivoire

Gooré Bi Tra
Laboratory of Mathematics and New Information Technologies, National Polytechnic Institute
Yamoussoukro, PO Box 1093, Côte d'Ivoire

BROU Konan Marcelin
Laboratory of Mathematics and New Information Technologies, National Polytechnic Institute
Yamoussoukro, PO Box 1093, Côte d'Ivoire

Souleymane OUMTANAGA
Laboratory of Computer Science and Telecommunications, National Polytechnic Institute
Abidjan, PO Box 475, Côte d'Ivoire

***Abstract*—The purpose of this paper is to provide a decision support tool based on a mathematical model and an algorithm that can help in the assessment of the level of vulnerability of children in Côte d'Ivoire. So, this study was conducted in three phases, the first one includes the settlement of a data warehouse. Then the second involves the application of probabilistic model. The final phase deals with the classification of children considered vulnerable in descending order from the most to the least vulnerable. The purpose of this classification is to better manage the resources of donors to support vulnerable children. This work is part of the activities of UMRI The resilience of Côte d'Ivoire. This is to propose mathematical and computational tools to facilitate the work of the Centre for social resilience. The use of the context of children made vulnerable due to crises or diseases is an example of practical application of our social resilience model**

***Keywords*—Crisis; Children; XML data warehouse; data mining; scheduling; Resilience; snowflake pattern; vulnerability level; probabilistic model**

I. Introduction

Today, while data abound, resources are dwindling, it is therefore necessary to have a tools for decision support based on computer and information technologies. These are based on data warehouses which are nothing but conventional databases, the only difference being that the underlying model does not use the entity-base relationship formalism. One of the models used to build data warehouses is the snowflake schema. Regarding the activities of social resilience, we necessarily need data storage tools that will be used to support the analysis of data collected in the data warehouse. To offer a dimensional model suitable for the settlement of a data warehouse dedicated to children's resilience. Just as a modeling approach giving an example of the exploitation of data. Thus the use of Bayesian networks technology and dimensional modeling to analyze the level of vulnerability of children in Côte d'Ivoire will be discussed later in this article in order to ease the decision-making of facilitators for the support of children considered vulnerable.

II. Review of Literature on Social Resilience

The literature on the concept is very diverse. Etymologically, it means resisting and bounce in front of a significant and persistent adversity. No consensus was reached on the definition of the concept as those proposed are linked to cultural considerations and therefore vary according to societies and also from one period to another. Resilience according to some [1] researchers, is a set of personal characters of the individual (or group of individuals), a process and an outcome. It is part of a learning process, self-determination through which the person interprets the meaning of a situation of adversity positively and reorient the direction of his life to pursue its development while strengthening its protective personal or environmental factors. However, with the situation of adversity as new self-organizer of the individual. However, all of them have in common the ability to bounce back from a shock and adapt to change [2] [3]. In this perspective, any measures of resilience require the observation of the demonstration of several dimensions. The definition used in this study is that of Tisseron [4] "Resilience is both the ability to withstand the trauma and to recover after it.

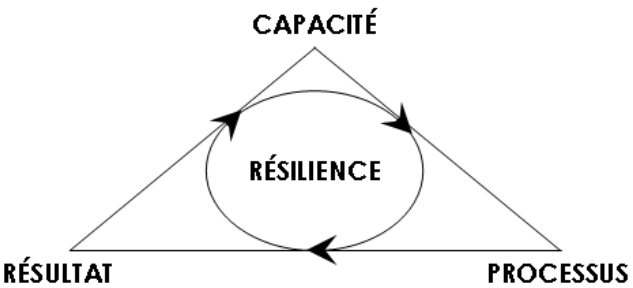


Fig. 1. Model illustrating the resilience [9]

III. Dimensional Modeling of Vulnerable Children Judged Resilience Data

In Côte d'Ivoire, the issue of vulnerable children support has led Organizations to put into place monitoring systems and feedback devices without forgetting their care taking. Very often support policies emanate from international programs. In That way, dimensional modeling aims to present data in a standardized form to facilitate intuitive querying of data. The subject for analysis is in the center of the model (fact table) and the relevant area to be analyze are attached to it (dimensions).The following table presents the different dimensions included in the monitoring and evaluation system of vulnerable children by NGOs led by HOPE-CI and funded by an American government led program called PEFAR section Cote d'Ivoire.

TABLE I. Monitoring and Evaluation Dimensions of Vulnerable Children from CSI1

DIMENSIONS	ATTRIBUTES	ARRANGEMENTS
Food and nutrition	Food Safety	Although, Average, Poor; Very bad
	Growth and Nutrition	Although, Average, Poor; Very bad
Education and Performance	Education	Although, Average, Poor; Very bad
	Performance	Although, Average, Poor; Very bad
Housing and care	housing	Although, Average, Poor; Very bad
	care	Although, Average, Poor; Very bad
Health	Health	Although, Average, Poor; Very bad
	Health services	Although, Average, Poor; Very bad
psychosocial	Emotion	Although, Average, Poor; Very bad
	Social behavior	Although, Average, Poor; Very bad
Protection	Abuse and Exploitation	Although, Average, Poor; Very bad
	Juridic protection	Although, Average, Poor; Very bad

The table shows the structure of monitoring and evaluation information allowing to assess the level of vulnerability and resilience of vulnerable children. Starting from this, we can construct an improved dimensional model, adapted to the implementation of data warehouse and the monitoring and assessment of the level of children's vulnerability. In fact, dimensional modeling, and notably the snowflake schema, is well known for its effectiveness in developing solutions in decision making. It is easily exploitable for developing reporting applications and dashboards. From a conceptual standpoint dimensional modeling is related to the concepts of fact and dimension:

Subject = (fact + dimension)

The snowflake schema derived from Table 1 is given by the following figure:

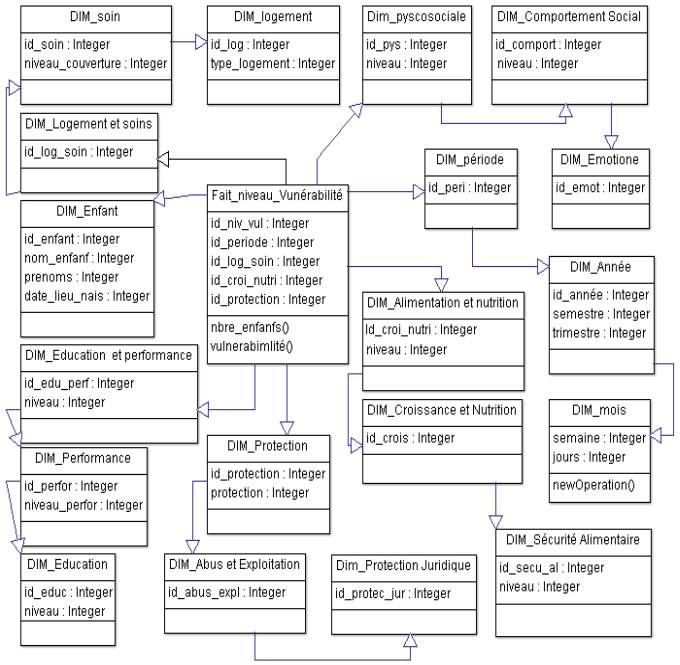


Fig. 2. The dimensional model snowflake vulnerable children

To succeed in the implementation of the snowflake schema, the schema of the basis must be the same as the one currently used (spreadsheet) be converted into the format of the data warehouse by ETL program. In addition to this, the quality of stored data should be considered in the practical implementation of the data warehouse. For good data quality, it is still possible to extract useful data for a multidimensional search, achieving, by the way, all the necessary corrections for their operation. The interest of having a data warehouse is to be able to regularly examine the resilience of vulnerable children in order to optimize the financial resources deployed by donors. [5] Several non-governmental organizations and public health facilities are overwhelmed with data but do not have the information they need to make good decisions. Knowing that they have all the data in warehouses can help these organizations optimize their decision making. Therefore, the research on vulnerable children, the analytical method requires methods that are easy to update and that provide a simple and efficient simulation approach. In this context, Bayesian Networks technology is the right approach because of its character both qualitative (these algorithms) and quantitative (embedded graphs). [6] The following figure shows the functional architecture of the core application of the decision-making system to create.

[1] CSI: Children Status Index record (*Index of Child Status Rating*)

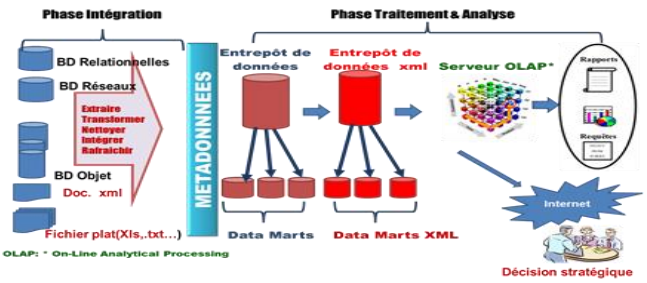


Fig. 3. Functional architecture of XML DW

IV. Bayesian Network Resilience Modeling

Also known as probabilistic expert systems, Bayesian Networks are tools of knowledge representation and automated reasoning on that knowledge. They were introduced by Judea Pearl in the 1980s and are found to be powerful useful tools for representing uncertain knowledge and reasoning from incomplete information. Bayesian Networks are simulation tools for observing the behavior of a complex system in contexts and conditions that are not necessarily accessible to experimentation. Technically, Bayesian networks are graphical models combining graph theory and probability theory. The following diagram (2) shows an example of Bayesian Network [8]:

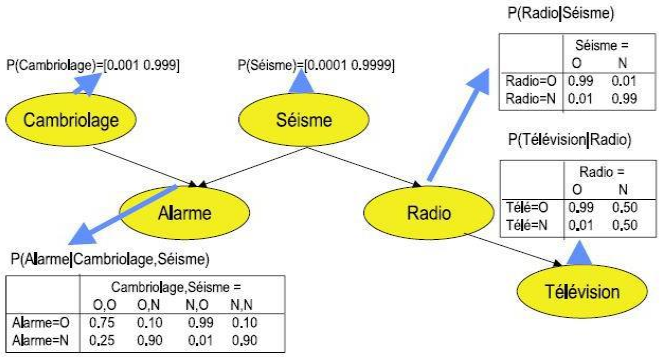


Fig. 4. Example of Bayesian Network

The above diagram is taken from a tutorial presented at the 8th scientific meeting dedicated to Knowledge Discovery in Data from (Philippe Leray). This Bayesian Network [09] models the process of triggering a security alarm in an environment frequently subjected to earthquakes. These earthquakes affect radio facilities upon which the television infrastructure is built. As shown in figure (Fig 2), a Bayesian network is a directed graph in which nodes represent the variables and the arcs symbolize the dependency relationships between these variables. Each node has a conditional probabilities table that is a model of beliefs in the occurrence of a particular case when we are in such a condition.

V. Probabilistic Modeling of the Resilience of Vulnerable Children

According to the review of literature, the Bayesian Networks were introduced by Judea Pearl in the 1980s and turned out to be powerful and useful tools for representing uncertain knowledge and reasoning from incomplete information. Representational knowledge to and automated reasoning about knowledge. Bayesian Networks are simulation tools to observe the behavior of a complex system in contexts and conditions that are not necessarily accessible to experimentation. Technically, Bayesian networks are graphical models combining graph and probability theory. In the framework of the modeling of the process of identification of best actions of resilience process, the graph of Bayesian networks reflects the identified actions and decision variables thereon**.** The graph of figure4 depends on the computer model that helped create the data warehouse. [10] We will use the graphic function of the software GeNIes[2], Which is a powerful simulation Bayesian network tool and especially free, to model the process of understanding the level of child vulnerabilities. The following chart provides the structure of relationships between the different attributes:

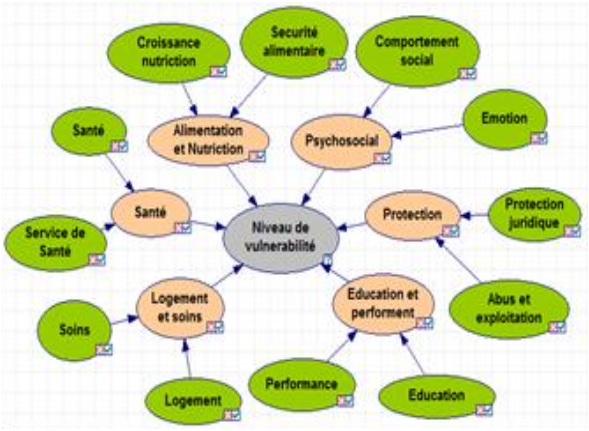


Fig. 5. Structure of Bayesian Network developed from the flake dimensional model

The above graph presenting the structure is an intuitive representation of the dimensions of resilience that governs the process of detection of children's vulnerability level. This representation was obtained after the introduction of the dimensions obtained from the dimensional model proposed flake. This graph of dependencies of Figure 4, is the qualitative part of the corresponding Bayesian network model. Although the conditional probabilities can be provided by experts in the field, the fact of having already structured data provides more accurate estimates of these probabilities.

VI. Association Gate Design for Bearing-Only Target Tracking

In practice there are many free software developed for their implementations. This reflects their importance in the field of machine learning [11]. For this project we have decided to implement the data warehouse part of our contribution on a free software Pentaho which integrate both tools, ETL, OLAP and reporting tools with a possibility to use data mining techniques for the analysis. Once the parameters are estimated by the application, the use of Bayesian Networks is to simulate the effects, of a number of choices on all other actions and variables included in the model developed. It includes a large number of learning algorithms as well as the

[2] Website : ***https://dslpitt.org/genie/***

parameters of the structure from the data [12]. It also has a friendly interface and can easily be used by non-specialists in modeling, including policymakers. For the simulation phase we will classify through an algorithms the degree of vulnerability of children considered vulnerable in the first algorithm with Bayesian networks. For the practical phase we are going from the base end of the first simulation data with GeNies3 Use the R language to run our algorithm on a sample of 51 children (Table 3). After that we will analyze the new results and finally we would draw a conclusion. Here is the scale of values transcribed in the table below.

TABLE II. INCREASING SCALE OF THE LEVEL OF VULNERABILITY

status / children	very vulnerable	vulnerable	Acceptable	resilient
Intervals of the scores	Rubbish > 50%	Bad> 50%	Average> 50%	Good> 50%
increasing scale	1	2	3	4

TABLE III. EXTRACTED FROM THE DATABASE R

	N..Ordre	NomE	NV
1	1	COULIBALY AWA	1
2	2	KONE TIEKOURA	3
3	3	KOUAKOU MAURICE	3
4	4	N'GUESAN GERARD	2
5	5	KOFFI YAO SERGE	1
6	6	KOBENAN ALI	1
7	7	OUTTARA NOUHO	2
8	8	KANGAH KOFFI	3
9	9	ADEPAUD HERVE	1
10	10	COULIBALY AYA	3
11	11	KONE TIEKOURA	3
12	12	KOUAKOU EDIE	2
13	13	N'GUESAN MEDARD	2
14	14	YAO KONAN ALBERT	2
15	15	KOBENAN ALI	1
16	16	OUTTARA JEROME	3
17	17	KANGAH KOFFI	4
18	18	KARNAN JEROME	1
19	19	COULIBALY AWA	2
20	20	KONE TIEKOURA	2
21	21	KOUAKOU MAURICE	2
22	22	KOUASSI ADLES	2
23	23	YAO SERGE	2
24	24	KOBENAN ALI	1
25	25	OUTTARA DRISSA	2
26	26	KANGAH KOFFI	3
27	27	KOUYATE BENGALI	1
28	28	COULIBALY ABIBA	2
29	29	KONE TIEMOMAN	1
30	30	KOUAKOU MAURICE	3
31	31	N'GUESAN GERARD	2
32	32	YAO YAO SERGE	3
33	33	KOBENAN ALI	2
34	34	OUTTARA NOUHO	1
35	35	KANGAH KOFFI	2
36	36	KONE MALICK	3
37	37	COULIBALY AWA	2
38	38	KONE TIEKOURA	3
39	39	KOUASSI MADISON	3
40	40	N'GUESAN GERARD	1
41	41	YAO SERGE	2
42	42	KOBENAN ALI	2
43	43	BGAME SILVAIN	2
44	44	KANGAH KOFFI	2
45	45	KONATE NAVIGUE	2
46	46	COULIBALY MAWA	1
47	47	KONATE TIEKOURA	2
48	48	KOUAKOU MICHEL	2
49	49	KONAN GERARD	3
50	50	YAO KOUAKOU	3
51	51	TOURE ALI	3

ALGO matrix_ordonnancement
Entrance
popu: child list
beginning

browse popu;
each iteration comparing the value
vulnerability to the previous item
that of the next item;
Yes *Previous item value> then next item value*
swap with previous item next item
Repeat *browse popu until there*
is more permutation
display ranked list ()
END

Fig. 6. Pseudo scheduling algorithm level of vulnerability

Fig. 7. List of children by level of vulnerability R

The TABLE III shows an excerpt of data sorted by the proposed algorithm Figure5. This algorithm also provides a graphical visualization of single child vulnerability of the state being in the database grouped by degree of vulnerability as shown in figure 6:

[3] Website : ***https://dslpitt.org/genie/***

TABLE IV. EXTRACTED FROM THE DATABASE BASED ON SLICES

Children age group	Number of children assessed	Food Safety			
		1	2	3	4
Aged 0 to 23 months	1	0	0	1	0
Boys 2 to 4 years	8	0	2	6	0
Boys aged 5 to 9 years	8	0	0	8	0
Boys 10 to 14 years	5	0	1	4	0
Boys 15 to 17 years old	2	0	1	1	0
total Boys	24	0	4	20	0
Boy total control			24		
Girls aged 0 to 23 months	0	0	0	0	0
Daughters of 2 to 4 years	5	0	2	3	0
Girls 5 to 9 years	8	0	3	5	0
Girls aged 10 to 14	8	0	2	6	0
Girls 15 to 17 years of age	6	0	0	6	0
total Girls	27	0	7	20	0
Girls total control			27		
TOTAL Boys & Girls	51	0	11	40	0
Total 0 to 23 months	1	0	0	1	0
Total of 2 to 4 years	13	0	4	9	0
Total of 5 to 9 years	16	0	3	13	0
Total of 10 to 14 years	13	0	3	10	0
Total 15 to 17 years old	8	0	1	7	0

Legendre:

1: Highly Vulnerable, 2: Vulnerable, 3: Good 4: Resilient

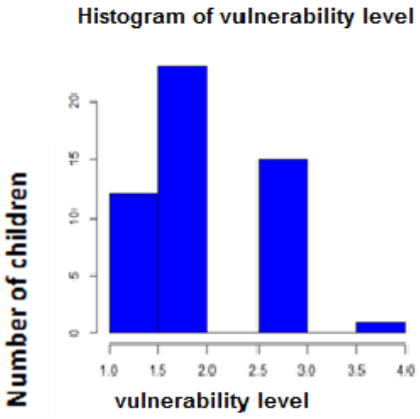


Fig. 8. Overall level of vulnerability of children-wide growth

For an interpretation we will consolidate the results on the level of vulnerability of children per age group. They will be classified according to the dimensions of social resilience. Here are the different age ranges ([0 to 23 months], [2-4 years], [5-9 years], [10 to 14], [15 to 17])).

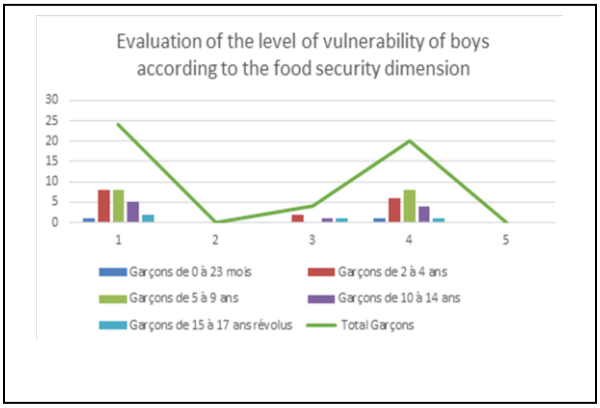


Fig. 9. Simulation result of the level of vulnerability of boys to food security dimension

These figures illustrate an example of evaluation according to food security dimension; it will be the same for the evaluation of each of the different dimensions in the end a conclusion will be drawn.

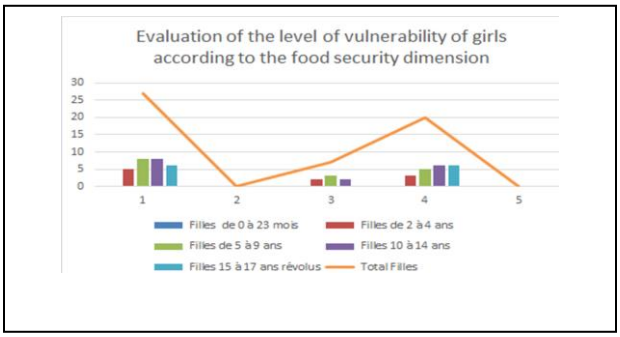


Fig. 10. Overall Level of girls' vulnerability to food security dimension by age group

The pace and curves of the graphs show the level of vulnerability among both young boys and girls is almost identical. Besides, considering the age range of the children, between [5-10] are less vulnerable than others see Figures 8 and 9.

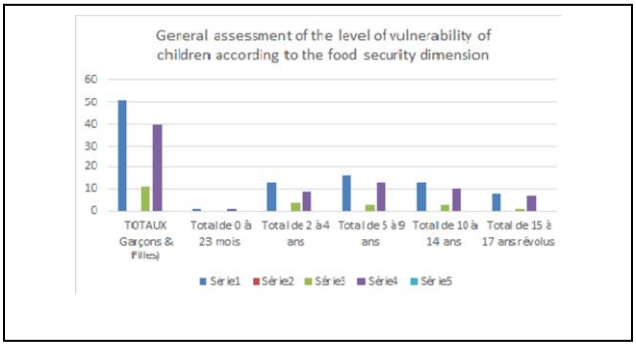


Fig. 11. Overall level of vulnerability of children (girls and boys) for food security dimension

In a word, it appears from this experience that whatever the sex of the children, the most vulnerable age group is the

range between [2-10] years for food security dimension. This could be explained by the fact that children at this age begin a growth phase, it then takes more resources to the parents to diversify their diet. However we usually have to do with orphans who live with guardians sometimes in households considered vulnerable themselves [12].

VII. CONCLUSION

The care for children in difficulty in order to reduce the level of poverty in African states and even in developing countries is a major issue. Non-governmental organizations and international organizations such as PUNUD, WHO and UNAIDS regularly develop and fund projects for their support. But, the resources are limited in face of the needs expressed. So, there is a need to use mathematical and computational tools in order to optimized management of these resources, for the storage of information collected in a data warehouse will significantly improve not only the management of these data, but also their use for purposes of decision support, particularly in the understanding of the process of the resilience of vulnerable children .In the framework of the analysis of resilience in general and that of vulnerable children in particular, the Bayesian networks are particularly appropriate owing to the fact that they are adapted to situations where one is confronted with incomplete, inaccurate and uncertain data . The use of simulation GeNies[4] helps us just to know the status of a child. To know the children whose situation requires an imminent and total support? Also, on top of the knowledge of their status, they must be determined and classified according to their level of vulnerability. This will enable organizations to make optimal management of resources allocated by donors. The major advantage that our contribution will make is that it assesses the dimensions and function separately based on the sex Children. This helps understand the importance and the impact of a dimension on the whole. As regards the results indicated on the figure above in all there are not any vulnerable or resilient children.

[4] Website : ***https://dslpitt.org/genie/***

REFERENCES

[1] Bahadur, AV, Ibrahim, M. & Tanner, T. 2010. The resilience renaissance? Unpacking of Resilience for Tackling climate change and disasters. Strengthening Climate Resilience Discussion Paper 1. Brighton, UK, Institute of Development Studies, University of Sussex. Available at: http://community.eldis.org/.59e0d267/resilience-renaissance.pdf. Accessed August 10, 2012.

[2] Anaut M. (2003) Resilience: Overcoming trauma, Ed. Nathan University.J. Clerk Maxwell, A Treatise on Electricity and Magnetism, 3rd ed., vol. 2. Oxford: Clarendon, 1892, pp.68–73.

[3] Anaut, M. (2002). Trauma, and Vulnerability Resilience in Child Welfare. Connections magazine, Eres, Volume 77, 101-118.

[4] TISSERON, S. Virtual adolescence: self-destruction or self-help? Neuropsychiatry of childhood and adolescence, 2007, Vol. 55, No. 5, p. 264-268.

[5] Achiépo Odilon Mr. Yapo (2015). "Plspm. formula: Formula Based PLS Path Modeling". R Package version 1.0.1, December 2015. https://cran.r-project.org/package=plspm.formula.

[6] Narchi Saud, Joelle. Factors associated with resilience of family caregivers of an elderly parent with loss of independence at home in Lebanon. 2014.

[7] TIEKOURA, Coulibaly Kpinna and BOKO Aka. SimCT: A measure of semantic similarity to adapté hierarchies of concepts. International Journal of Computer Science and Information Security, 2016, vol. 14, No

[8] SAHA Bernard Kouassi, ACHIEPO Odilon Mr. Yapo, BROU Konan Marcelin, et al. Storage and Bayesian modeling of data on the social resilience: Case of Orphans and Vulnerable Children (OVCs) in Ivory Coast. International Journal of Computer Science Issues (IJCSI), 2015, vol. 12, No. 4, p. 137.

[9] R. Kimball, R. Mertz, "The Data Webhouse: Building the Web-enabled Data Warehouse", John Wiley & Sons, 2000;

[10] Von Overbeck Ottino S. (2002) Mental Health in the city. Resilience in psychotherapeutic work: what psychological support for children exposed to mob violence? in psychiatric Papers, No. 29, BDSP.

[11] SAHA Bernard Kouassi, BROU Konan marcelin, Babri Michel, et al. Classification of Households in after-traumatic shock, with the aid of Bayesian Networks: example of the post-electoral crisis in Côte d'Ivoire. International Journal of Computer Science and Information Security (IJCSIS), 2016, vol. 14, No. 7, p. 201.

[12] Melaine Achiépo Odilon Yapo, ABIDJAN, Ivory Coast, PATRICE, Mensah Edoeté, et al. Resilometrics: Principles of the Discipline and the Resilience Measure Models. International Journal of Computer Science and Information Security, 2016, vol. 14, No. 5, p. 162

Software-Defined Networks (SDNs) and Internet of Things (IoTs): A Qualitative Prediction for 2020

Sahrish Khan Tayyaba
Department of Computer Science,
COMSATS Institute of Information Technology,
Islamabad, Pakistan

Naila Sher Afzal Khan
University of Management Sciences and Information Technology,
Kotli, AJK

Wajeeha Naeem
Department of Computer Science,
COMSATS Institute of Information Technology
Islamabad, Pakistan

Munam Ali Shah
Department of Computer Science,
COMSATS Institute of Information Technology
Islamabad, Pakistan

Yousra Asim
Department of Computer Science,
COMSATS Institute of Information Technology
Islamabad, Pakistan

Muhammad Kamran
Department of Distance Continuing & Computer Education,
University of Sindh,
Hyderabad, Pakistan

***Abstract*—The Internet of Things (IoT) is imminent technology grabbing industries and research attention with a fast stride. Currently, more than 15 billion devices are connected to the Internet and this number is expected to reach up to 50 billion by 2020. The data generated by these IoT devices are immensely high, creating resource allocation, flow management and security jeopardises in the IoT network. Programmability and centralised control are considered an alternative solution to address IoT issues. On the other hand, a Software Define Network (SDN) provides a centralised and programmable control and management for the underlying network without changing existing network architecture. This paper surveys the state of the art on the IoT integration with the SDN. A comprehensive review and the generalised solutions over the period 2010-2016 is presented for the different communication domains. Furthermore, a critical review of the IoT and the SDN technologies, current trends in research and the futuristic contributing factors form part of the paper. The comparative analysis of the existing solutions of SDN based IoT implementation provides an easy and concise view of the emerging trends. Lastly, the paper predicts the future and presents a qualitative view of the world in 2020.**

***Keywords*—SDN; IoT; Integration of SDN-IoT; WSN; LTE; M2M communication; NFV**

I. INTRODUCTION

The emergence of new technologies and communication networks offer new connectivity scenarios among every physical object. Machine-to-Machine (M2M), Device-to-Device (D2D), Vehicle-to-Vehicle (V2V), wireless sensor network, actuators, smartphone, embedded devices and even connections among infrastructures are developing new connectivity scenarios. Moreover, these devices will be allegedly connected to the Internet and will ultimately create a heterogeneous system of interconnected objects; called the Internet of Thing (IoT), and in broader sense Internet of Everything (IoE) [1]. The IoT devices are generally sensor node, actuator, RFID tags and wireless communicating devices connected to the Internet in a smart environment. The IoT devices are capable of observing, analysing and taking intelligent decisions based on collected information from the surroundings and manipulation of the underlying network. The IoT devices are deployed according to the customised task with specific applications; forming a domain specific IoTs network. This domain specific applications and service attribute a horizontal view of the IoT network such as appliances and applications for smart home management, smart health care unit implanted on the body or wearable sensors for health monitoring. The domain-based services can leverage the benefits of pervasive and ubiquitous computing through the independent services horizontal platform.

With the immense increase in IoT devices huge amount of data is generated and collected which impede monitoring, management, controlling and securing IoT devices in a heterogeneous network and become a critical issue for researchers and developers. Traditional network does not completely support heterogeneity, which limits IoT benefits full realisation. In addition, the services demand and customers require fast development and deployment that is still an issue in a traditional network. The innovation in the legacy network is very slow due to the proprietary nature of devices. Therefore, a change in the traditional network infrastructure and devices is mandatory to realise full benefits of IoTs. IoT can leverage full benefits from the integrated architecture of such technologies. The most attracted technologies in this domain are Software Defined Networking (SDN), and Network Function Virtualization (NFV).

SDN is an emerging technology that can meet the need of current IoT requirements of heterogeneity and flexibility. It provides a centralised control and global view of the whole network. SDN decouple the control functionality from the forwarding plane and program network service sitting above the controller (control Plane). The centralised management facilitates optimisation and configuration of a network in an efficient and automated manner and provides interoperability among heterogeneous IoT network. This control plane centralization can provide a secure architecture for IoT

network, e.g., smart home security applications prevent unauthorised user access of the smart appliance etc. IoT is growing with a very fast stride that new trends and technologies, protocols, architecture, management, and security solutions in the context of IoT are formulated within a short period. There is a research gap in addressing the IoT integration with different networking solutions especially, leveraging the benefits of SDN.

In this paper, we highlight different studies which provide SDN based solutions for IoT technologies. We survey the literature over the period 2010-2016, by focusing the attention on different aspects of the IoT merger with the SDN. The organisation of this paper is as follows. Section II provides some background of the IoT and the SDN and architecture of two contributing domains, i.e., SDN and the IoT and the protocols for the SDN. In section III, a comprehensive literature is provided for the existing solution of the SDN and the IoT integration. Section IV provides a detailed review of the existing solution, providing a comparative analytics of the existing integration solutions. In section V, market and research trends and a qualitative prediction for 2020 are given. Section VI concludes the study.

II. Background Related Studies

A. Background

The use of computing devices and communication technologies are growing exponentially with the decline in cost and size of hardware and software. Vendors and organisations are digging new domain in search of finding new ways of flexible computing and communication. IoT and SDNs are two complete different communication and network domain whose merger is seeking for benefiting human kinds and developing smart systems. As the IoT implementation expectancy exceeds the limits of traditional network e.g., Virtual Private Network (VPN), the SDN promise to hold the traditional network with new service demands. At this stage, technology shift is highly intention grabbing a task from the researchers and developers in industries and organisations. The two domains and their architecture are totally dissimilar. In this section, an architectural detail of both domains is presented to grab the underlying functionality for the merging of IoT in SDN.

B. SDN Architecture and protocol

In a traditional network, the devices and the equipment are usually proprietary entities, are physically distributed and control function is hard-coded. The network operator has to do configuration of the individual network device as per service layer agreements (SLAs) and cannot be programmed otherwise. The complexity increases due to the vertical integration of network architecture. The control plane and the data plane are bundled inside the networking devices, reducing flexibility and hindering innovation and evolution of the networking infrastructure. Any change in the network is expensive in term of time, and cost. The cost comes in term of capital expenditure (CAPEX) and operational expenditure (OPEX) [2]. For example, the transition from IPv4 to IPv6, started more than a decade ago and still largely incomplete, bears witness to this challenge, while in fact, IPv6 represented merely a protocol update. To overcome the existing architecture, SDN is considering as the best alternate.

In SDN, the control plane is decoupled from forwarding plane and communication between two planes is done through using Southbound and Northbound APIs. SDN is basically layer architecture consists of three layers 1). Device layer or data plane 2). Control plane and 3). Application layer. The customer needs are abstracted over application layer which is communicated to the controller via Northbound APIs e.g., RESTfull API. The control layer or controller is centralised part of the SDN network and act as a brain of the network. The controller manages the whole network and possesses a global view of the network. All applications/programs run above the controller. Many controllers are in the market from its inception such as ONOS, Open daylight, Floodlight, NOX [3], POX, Trema etc. SDN controller define rule for the incoming flows from the data plane. The controller communicates with the devices in the data plane via Southbound APIs, most common and recognised is OpenFlow (OF). The layered architecture of SDN is shown in Figure 1

SDN do not increase the performance of the network rather it provides flexibility in network configuration and resource management. On the contrary, SDN can lead to performance degradation in case of providing high level of abstraction

1) SDN architecture

SDN is a layered architecture, consisting of three basic layers; application/services layer, controller layer (control plane), and data plane layer called forwarding layer consisting of forwarding devices. These SDN layers communicate with each other via open APIs called Northbound Interface (NI) API and Southbound Interface (SI) API [5]. To identify the different elements of an SDN as clearly as possible, we now present the essential terminology used throughout this work

a) SDN architectural components

SDN is a layered architecture, consisting of three basic layers; application/services layer, a controller layer, and data plane layer called forwarding layer consisting of forwarding devices. These SDN layers communicate with each other via open APIs called Northbound Interface (NI) API and Southbound Interface (SI) API [5].

SDN layered components are described to

- *Application layer (AP):* The application plane also called management plane consist of applications that leverage the functions offered by the NI to implement network control and operation logic. Essentially, a management application defines the policies, which are ultimately translated to southbound-specific instructions that program the behaviour of the forwarding devices.

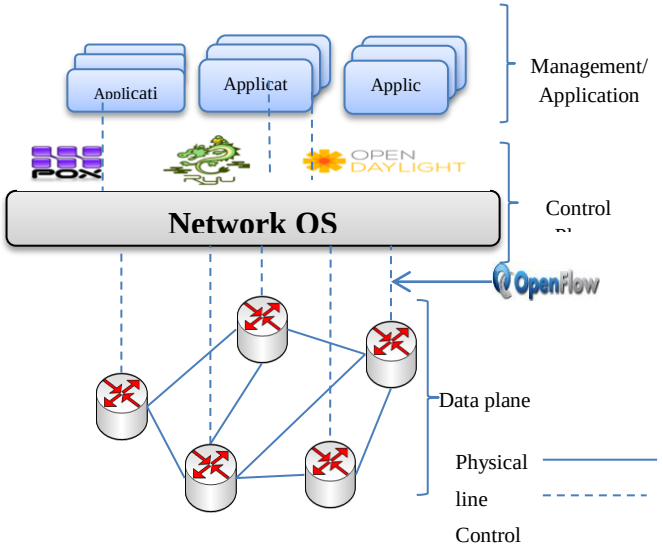


Fig. 1. SDN Architecture

- *Northbound Interface (NI):* The interaction between application AP and control plane is provided through NI. The Network Operating System (NOS) facilitate application developers to coordinate through these NI APIs. Typically, an NI APIs abstracts the low-level instruction sets and implementation of forwarding devices. So far NI APIs is not well studied. Generally, RESTFull APIs are used as an interface between applications and control plane.
- *Control Plane (CP):* Control plane is the decoupled entity from the distributed forwarding devices and logically centralised on a server. CP programs the forwarding devices through southbound interfaces. CP defines rules/instruction set for forwarding devices hence control plane is the ‘‘network brain’’ and all control logic rests in the applications and controllers, which form the control plane. Many SDN controllers are available in the market such as NOX[3], OpenDaylight[5], Ryu[6].
- *Southbound Interface (SI):* Southbound interfaces provide a communication protocol between CP and forwarding device though the SI instruction set. Well established SI protocol help controller in programming forwarding devices and formalise rules for interaction between the two planes (CP & DP). Some examples are OpenFlow [7], Forwarding and Control Elements (ForCES) [8] , Protocol-oblivious forwarding (POF) [9].
- *Forwarding Devices (FD):* Network core devices either software based or hardware based performs fundamental network operations. The forwarding devices act on the basis of rules/instruction set provided by CP/controller on the incoming flow/packets (e.g., forward, drop, rewrite some header). These instructions are defined by southbound interfaces such as OpenFlow [7], ForCES [8] and are installed in the forwarding devices by the SDN controllers implementing the southbound protocols.
- *Data Plane (DP)/Forwarding Plane:* Forwarding devices (routers, switches, gateways etc.) are interconnected through a physical medium such as wireless radio channels or wired cables. And defined a physical interconnection within a network

SDN has many applications in other networks such as in management, configuration and reconfiguration of the network in a flexible manner. SDNs provide a fine-grained control with high quality of services. The SDN controller flexibly manages the flow forwarding state in the data plane (router & switches) by having a global view of the network. SDN controller provides programmability for the data plane. Controller is logically centralised entity but physically distributed [10]. SDN is believed to provide its user with a separate networking slice by utilising the concept of virtualization. NFV is considered as a complementary technology for SDN. SDN utilised the virtual view of the network status and provide different applications based on this virtualized view. NFV can be implemented as an application above the CP. Network functions can be virtualized in NFV. The next generation network architecture is quite dependent on such technologies which can facilitate high data transmission, spectral efficiency, resource allocation and network management for fulfilling growing need of the customer demands. One solution to such demand is the programmability of the network and dynamic allocation of resources, which can be provided by network virtualization. In virtualization, user specific network is called slice, which provides new values to user requirements and applications. In the next section, we will highlight the detailed architecture of IoT network, which is again layered architecture of connecting the physical object with the Internet.

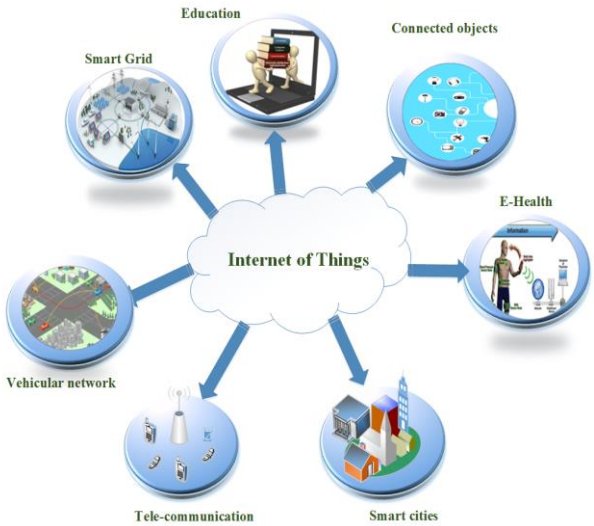


Fig. 2. Overall IoT scenario

2) IoT architecture

We are living in the era of connected objects where devices can communicate with the physical world and capable

of taking decisions due to the data analytics. The main factor behind this swift shift is the advancement in the microelectronics, telecommunication networks, and use of Radio Frequency Identification (RFID) tags attached to the physical objects. When these objects are connected to the Internet, they form a network of interconnected objects called IoT. The IoT is simply the point in time when more things or objects are connected to the Internet than people. [11]. As the boundaries of connected objects are not limited to certain technology, diverse ranges of objects connect and communicate with each other using a different communication protocol, resulting in the heterogeneous network as visible in Fig. 2. IoT devices are used to sense, collect, process, infer, transmit, notify, manage, and store data. The IoT helps in building a smart environment. Few examples are home safety and management system, smart electricity monitoring in electricity grids, in-car system from road traffic monitoring to control function and safety measures in advances, health monitoring to smart building automatically controlled heating, venting, and air conditioning (HVAC) systems, security systems, disaster management, weather forecasting etc. are variant domain and provide a powerful control in handling daily life activities. There are billions of devices connected to the heterogeneous network. These entire domains have different architectural details as per the specified functional requirement and still not converged on are not converged on a single reference model [12], which add complexity in the heterogeneity of a network. However, the general architecture of IoT is shown in the Fig. 3

a) IoT architectural components

For any network, layered architecture ensures flexibility and capability of invocation of new services in the network, IoT architecture follows layered architecture. Due to varying IoT domain, architecture and contributing components are not converged however most successful IoT architecture is IoT-A [13]. Many other IoT architecture models are also in the market but most common is "four-layer architecture"

- *Perception layer:* Perception layer is physical object layer consisting of sensors, actuator, RFIDs, mobile devices, motes, blue tooth etc. This layer collects the data from the environment and transmits on the edge of the network i.e. gateway or sink.
- *Network layer:* This layer is responsible for transmitting data from physical objects to the gateway/edge of the network for further processing on the collected information. Different transmission technologies contribute to the heterogeneity of IoT such as ZigBee, blue tooth, Wi-Fi etc.
- *Application layer:* This layer deal with the application/services of the user demand by manipulating the information collected from the perception layer and processed in the processing system.

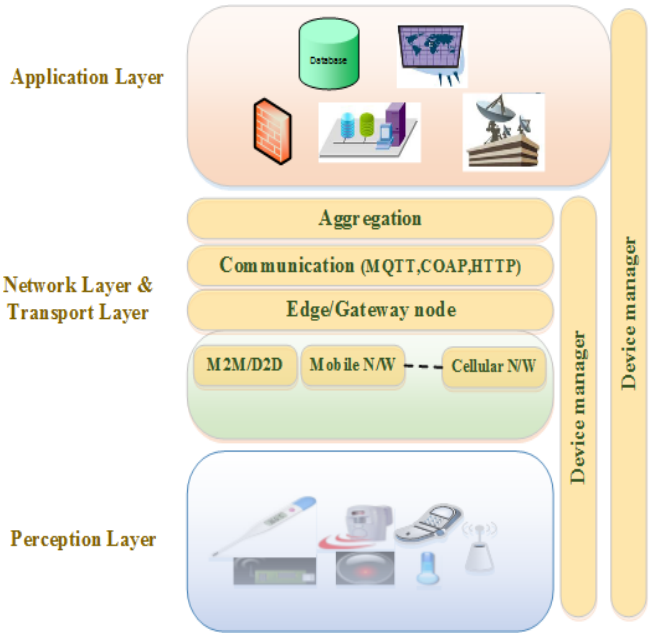


Fig. 3. IoT Architecture

- *Middleware layer:* Different IoT devices in a domain may be different but devices can interact with a compatible/same device. This layer translates the message of one service information without concern for the hardware detail. Middleware layer is associated with service management, addressing and naming of the requested service.

Beside these main layers, there are many components, which play important role in IoT information collection, processing and management. We define these components. *Edge services* component is responsible for delivering information through the Internet. These services may be domain name service, Content Delivery Network, firewall, load balancer etc. *Analytics services* component guide and automates the process of data analysis, discovery, and visualisation. The *Process management services* help in managing the workflow of the information processing and connects devices with their respective services. *Device identity* services identify a user registers service on a device. A*uthentication* service enables the authentication of a registered user with its associated service. *Service Oriented Architecture (SOA)* helps in providing architectural abstraction from the underlying detail and provides required services.

Initially, the Internet was distinctly established over TCP/IP suit and provided support for a large number of the connected computer. However, TCP/IP does not support heterogeneous network. Therefore, the TCP/IP is not suitable for IoTs. Hence, the heterogeneity of connected device in IoT environment is creating unprecedented complexity and functional diversity.

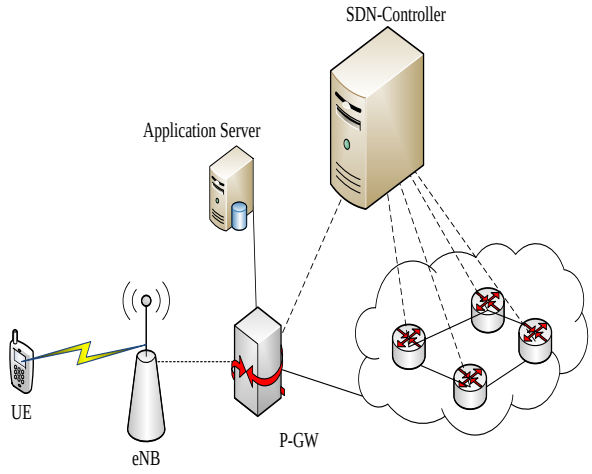


Fig. 4. SDN based LTE architecture

C. Related Studies

SDNs and IoTs are a hot topic and grabbing the attention of industry and market. Many comprehensive studies are done by the research communities to look into the detail perspective; implications, use cases, and technological demand in both domains. Kreutz et al. [2] present a comprehensive study on SDN. The authors provide a detail and all-inclusive on SDN, SDN evolution from programmable networks to SDN architecture, protocols, application, use case scenarios and future research trends etc. Nunes et al. in [3], discuss the past, present and future of programmable network based on SDN. SDN layered taxonomy is presented in [4]. Network innovation in the context of SDN using OpenFlow is dealt in [5].

An ample survey on IoT is presented by Al-Fuqaha et al. in [6], mentioning every domain of IoT, application, issues and scenarios. Similarly, Xu et al. [7] presents state-of-the-art on IoT. The future industrial perspective of IoT is presented in [8]. The study of IoT applications is done in [9]. The merger of IoT and SDN is also studied in many research articles as in [10] which presents the SDN and virtualization in IoT domain. However, a detailed survey on the integration of IoT in SDN requires attention from the research community.

III. Literature Review

Since SDN and IoT are in their infancy, still there are many problems and IoT use cases that are not completely realised. Even though IoT has a vast implementation in conventional routine creating scenarios with almost every network technology to extract information, bringing improvement in daily life and developing a smart ecosystem. In this section, we undergo an extensive review of the existing solution of IoTs based on SDN. Few of IoT implementation in the context of SDN based control and management is discussed below. SDN integration in current trends of IoT is a research question till yet. In this regard, many studies have been generated in the campuses and on the industrial level to get full advantage of programmability from SDN and Virtualization from NFV.

A. SDN Based D2D communication in LTE

Long Term Evaluation (LTE) is a communication standard evolved from Third Generation Partnership Project (3GPP) known as UMTS (Universal Mobile Telecommunication System) and introduces Multi Input multi-output (MIMO) to ensure high-speed data transmission at a higher data rate of 300Mbps peak downlink and 75 Mbps peak uplink [22]. It also provides connectivity of cellular network with the Internet using IP network equipment LTE support high data required services such as Voice over IP (VoIP), Video conferencing and multimedia streaming in a cellular network. It uses multiple radio access techniques and uses both Time Division Duplex (TDD) and FDD for downlink and uplink high data rate communication and improves spectrum efficiency. The working component of LTE are User Equipment (UE), eNodeB (access point), and EPC i.e. Evolved Packet Core. UE is actually a mobile used to link the user with the access network. The access network is an Evolved UMTS Terrestrial Radio Access. A general architecture for SDN based LTE is shown in Figure 4.

LTE, a major contributor in IoT, promise high data rate and low latency but despite these facts, LTE technologies encounter many issues of centralised control, Scalability and QoS challenges in the network. Centralised management and spectrum adjustment by operator minimises the automatic and dynamic control and management of the cellular network. In this context, several studies have been conducted based on the integration of LTE with SDN. In [14], LTE network reconfiguration is proposed using SDN based on D2D communication devices and ensure Quality of Experience (QoE) which is measured on the basis of Mean Opinion Score (MOS). Liu et al. proposed an algorithm for multi-tier LTE network reconfiguration for downlink and uplink based on a D2D communication protocol in case of congestion on the nearest eNBs. The parameters used to measure performance are download speed and waiting for the delay because of congestion in the adjacent eNBs. Savarese et al. in [15] proposed a Flexible approach for the reconfiguration and resource allocation in LTE environment when acting as IoT by observing context and connects various types of monitoring terminal devices and the Internet without human interaction. They use context-aware information and geophysical location for their proposed framework architecture for heterogeneous M2M devices over LTE/4G network with SDN controller and context-Aware Application (CAA) running over M2M server identifies the failure of certain eNB and informs SDN about the status. In CellSDN [16], Erran et al. proposed a cellular architecture based on SDN in which attribute-based policies are formulated for individual user in the LTE network and gain fine grain control over the network. CellSDN also proposed for SDN application for deep packet inspection by the local cell agent running in each switch. This local agent in CellSDN can increase scalability by reducing the excessive load on the controller.

As controller offload some of the measurement task to the local agent which can perform local control operations. In

[26], M. H. Kabir proposed cluster-based SDN controller architecture for a cellular network where the cellular area is divided into clusters controlled by a cluster controller where major functionalities are provided by SDN controller. Radio access related activities are controlled by SDN controller, which reduces the complexity in the based station, and load monitoring and session controlling is done through the controller's head in the clustered area. The cluster head controllers communicate with each other via controller services.

Legacy IoT mostly using IEEE802.15.4, ZigBee or 6LoWPAN (IPv6 over low power wireless Local Personal Area Network) protocol as communication protocol but 6LoWPAN protocol does not fulfil the required bandwidth need of IoT devices and do not create an efficient routing. An architecture framework is presented in [27], which uses SDN as the management platform for 6LoWPAN devices. SDN based Management Framework for IoT Devices is proposed in [28]. The author used SDN controller and three reference point for communication between different network entities and SDN Controller and focus on the transaction between M2M.

The communication between the private network and the public network is done through Network Address Translator (NAT), which exhaust when the number of devices increases in the network due to its centralised nature. Distributed NAT Traversal using SDN is used for managing IoT traffic by distributing the load on the SDN-enabled devices/switches and in result transmission delay is reduced [29]. The legacy NAT traversal scheme has many disadvantages as increased workload on the relay server, or inflexible P2P communication as required by IoTs, and performance degradation due packet modification and processing on each packet. But this is not an efficient way as the central SDN controller may also suffer the aforementioned problems in the NAT and NAT Traversal schemes also there is a single point of failure due to a centralised server.

Due to the huge amount of data produced by IoT devices and billions or devices are connected to IoT network, flow management is not an easy task. In cased of SDN based IoT architecture, where the controller is responsible for making flow rules, their installation at the gateway incur delay and degrade the performance of the network. This flow rule installation is hype when flows are installed reactively on demand. In [30], Bull et al. proposed pre-emptive flow rule installation by monitoring and learning the periodic behaviour of IoT network. In this proposed scheme, the flow rules are installed before the arrival of flow in the network by observing the flow history i.e. by learning switch techniques.

According to Cisco report, due to the immensely increasing IoT/mobile device and connection, Global mobile data traffic reached 3.7 Exabyte per month at the end of 2015, up from 2.1 Exabyte per month at the end of 2014 [31]. With such an immensely increased volume of data and traffic, the single centralised controller is not sufficient to handle generated traffic and flow management. An SDN centralised controller suffer from processing pressure as only a limited amount of flow can be processed by a single controller such as on NOX, around 30k flow request per second are processed. For this purpose, distributed controller solutions for SDN were proposed such as Onix, Open Network Operating System (ONOS), and DevoFlow etc. In IoT, this traffic flow management is important in term of heavy data especially video and audio streaming, multimedia contents and online gaming etc. which need extra care for defining management rules and policies.

In [17], the author presented a detailed review of the integration of Information Centric Network (ICN) in SDN. The integration of ICN and SDN over IoT devices is not an easy task because the significant solution for security and management is lacking in realising Sensing as a Service (SaaS) in SDN based IoT devices. A. El-Mougy proposed cloud application management in ICN using SDN CP. This integrated 5G/LTE network in SDN can also suffer from security risk of single point failure, minimization of transmission rate due to shared spectrum. In [18], Usman et al. proposed a hierarchal architecture for sensor IoT integration into 5G/LTE network using SDN domain controller. The architecture is monitored by central controller and other domain controller interacts with this central controller, this central controller dynamically allocates resource leveraging a D2D communication.

B. Middleware solution based on SDN

Different requirements for the two technologies are creating hazards for communication between IoT and SDNS. In [19], the interoperability of heterogeneous network in an IoT perspective is discussed and an architecture for communication between IoT and SDN environment is proposed using OMG Data Distributed Services model (OMG DDS) as middleware in which publisher/subscriber message are used for communication between different entities in a heterogeneous mode and provide scalability of a network. Similarly, CASSOWARY in [20], a provide a middleware architecture which helps in providing context aware communication in smart buildings using SDN based controller. CASSOWARY enables smart devices and SDN uses information to smartly handle the building HVAC system on the basis of distance and presence of activities or tenant in that environment.

In [21], Qin et al. enhanced the idea of Multi-network controller architecture for heterogeneous IoT network based on SDN controller for a multi-network environment such as network accessing Wi-Fi, WiMAX, LTE, ZigBee and another cellular network at the campus level and evaluated the performance by measuring delay, jitter and throughput. MINA is basically a middleware whose working principle is self-observing and adaptive, and manage the pervasive heterogeneous network. MINA takes advantage of SDN principle for flow matching and management. MINA follows SDN like layered architecture, which reduces the semantic gap between IoT and task definitions in a multi-network environment. The architecture is modelled using a Genetic algorithm and network calculus. Flow shares the same node resources and network is optimised for this resource sharing in this architecture.

WU et al. in [22], presents UbiFlow framework which provides the integration of the SDN and the IoT. UbiFlow proposed an efficient flow control and mobility management in urban multi-networks using SDN distributed controllers. In UbiFlow architecture, IoT network is partitioned into small network chunks/cluster in which each partition is controlled by a physically distributed SDN controller. The IoT devices in each partition may be connected to the different access point for different data requests. These distributed controllers coordinate to provide flow scheduling, mobility management, optimized access point selection in a consistent, reliable and scalable control order, and provide fault tolerance and load balancing for multi-network IoT. The per-device flow management and optimised access point selection are based on the multi-network capacity performed by the SDN controller, which partition the network using network calculus in the UbiFlow architecture. Ubiflow architecture is shown in Fig. 5

A representative summary of existing SDN based management Solutions for IoT given in survey are presented in Table 1.

C. SDN for wireless sensor based IoT devices

Wireless sensor network defines intercommunication of spatially distributed sensor node which is generally used as monitoring agent in the disaster areas, health care, environmental condition, industrial monitoring and earth sensing etc. The most common contributor in the IoTs is sensor nodes. Wireless Sensor Network (WSN) is deployed in different scenarios according to specific need e.g., sensor deployment for a volcanic study to deep-sea measurement, in the disaster area to dark forests reading throughout day and night. Many research articles articulated the role of wireless sensor nodes in smart ecosystem and contribution of telecommunication. However, tremendous growth in IoT devices/sensor node, application, collection and analytics on data need intelligence services and new paradigms. Our focus in this study is the integration of IoT component with SDN, so we collect reading based on WSN in the context of SDN. Mostly sensor node topology is a mesh topology or a peer-to-peer topology; management and control in constrained environment are always a vigorous research area.

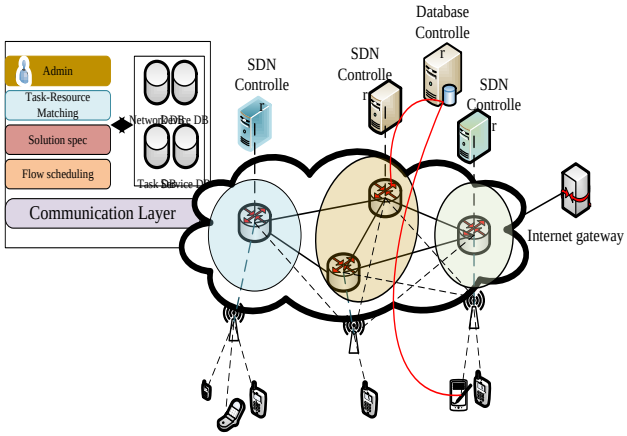


Fig. 5. UbiFlow architecture

In this context of mesh network connectivity, an interesting instigating architecture for a wireless mesh network (WMN) on the basis of OpenFlow was given by Delay et al. in [23]. In this paper, the author suggests the seamless mobility in the WMN by the use of OpenFlow. The KAUMesh test-bed allows the use of OpenFlow in WMN and provides an efficient and flexible mobility solution. In this solution, the mesh router is OpenFlow-enabled and contains multiple physical wireless cards. Multi-hop connectivity is achieved by using OLSR and data path uses local sockets to communicate with the control path component. Monitoring and Control Server (MCS) and NOX act as a controller interface and communication is done on a secure channel and handles all the flow rules. The association database contains a list of stations and the Mesh Access Point (MAPs). Connectivity graph can be obtained from gateways or may be from the QoS metrics. NOX handle routing task based on the information gathered from MCS. New rules are installed using topology database in the data plane. The associated station complies IEEE802.11standards and handover is done using IEEE802.21 standards. The OpenFlow protocol is used for setting up the flow tables, HTTP/XML for the communication between MCS and NOX and IEEE 802.21. This architecture is important as mesh connectivity play an important role in IoT scenario. However, the mobility model is suitable only for small scale while IoT implication is seen in larger context. The algorithm for the association of flow node and flow path in this architecture is undefined for MSN.

In the context of WSN management, few protocols are proposed such as SDN-WISE[24], Software-Defined Wireless Sensor Network Framework [25] and leverage SDN programmability in the WSNs. The architectural components of this approach consist of a Base Station (BS) and several sensor nodes. SDN controller operates on BS took a routing decision on the lieu of dumb sensor nodes. Sensor nodes contain flow table as in the SDN populated by controller.

WSN integration in SDN is seen in [41] with a three-layer architecture. It consists of master node/controller node, central node (OpenFlow enabled switch) and a normal node. The master node defines a routing policy for the normal node. The author et.al uses flowVisor as virtualization engine to make independent user slice in between controller and switch. Data forwarding is done by OpenFlow switch. The configuration and management are done by OpenFlow protocol, which also identifies the existing routing protocol, and work in congruence. The placement of central node is important. In this proposal, the distance is calculated based on cosine similarity formula. The central node locates in the physical centre of the cluster architecture, helps in maintaining network topology and help increasing network convergence. They name their architecture as SDWSN. Neighbouring node status is taken into account for the assigning role. Even though the author has verified their architecture through comparing its result with existing WSN protocol and find improvement in the result; the conceptual details are not very clear.

In [28], Miyazaki et al. proposed an architecture for reconfigurable WSN network on the basis of customer need

by using role injection and delivery mechanism. The role compiler generates scenarios which are injected through wireless communication. Field programmable array (FPGA) and a microcontroller unit (MCU) carry the change in the sensor nodes. The communication is done on the basis of sensor attribute. The role compiler is at the base station. The architecture consists of base stations (BSs), reconfigurable node and a server which contain role injection and delivery mechanism components. They name their architecture as SDWSN. Neighbouring node status is taken into account for an assigning role on the fly and sensor behaviour can be manipulated as per role description.

IoT devices with constraint resources are the main consideration while forming any architecture or protocol. The increased efficiency of sensor node communication is directly associated with energy management. Majority research focuses on sleep/active mode for energy restoration in the WSN. Wang et al. in [29] presented an SDN based algorithm "Energy Consumed uniformly Connected K-Neighborhood" (EC-CKN) called as SDN-ECCKN. In this architecture, a controller node calculates the overall energy of WSN. SDN-ECCKN helps in retaining energy of each node and minimises the broadcast messages from the individual node.

The multi-purpose sensor network is also addressed in [30]. Leontiadis et al. exploited NFV for sharing single infrastructure for many applications in a sensor network. They proposed a framework for multiple application scenarios on a common build infrastructure. Each node has an abstraction layer for a shared hardware which works on the overlay network and creates multiple virtual sensor networks (VNS). The bridge between application and hardware is written in TinyOS operating system. This is informally an idea of separating sensor node hardware plane for application oriented overlay VNS.

In [31], the author proposed an architecture for the integration of WSN with SDN controller. A local controller in each sensor node is responsible for MAC forwarding and some local routing decisions. A centralised controller is responsible for the long-term decision. In a sensor network, topology information collection is main challenge and different approaches are used for the information collection like packet trace, which contains detail information and Link Quality Estimation (LQE). The author suggested using lightweight LQE for collecting topology information, which can provide SDN controller with a global view of the network. This paper also proposed to take advantage of virtualization of SDN and change the object bytecode on the fly for commodity hardware. The SDN logical manipulation of virtualization and intelligent algorithms is used to get better IoT application and traffic analyser. Many of the implementation scenarios are also presented by the author.

Software Defined Wireless network (SDWN) [64], is an early effort for providing feasibility for the implementation of SDN for the wireless network. Costanzo.et al. presents architecture for Low Rate Personal Area Network (LR-PAN) management and flexible resource utilisation using SDN controller over the sink node. Sink node gathers topological information and coordinates this information to the controller, which defined rules/policies for better management. Each individual node computes RSSI factor for measuring network resource (local battery level and hop count). The rule or policies are defined by a controller implemented on a limited portion of incoming packets to safe space. However, this architecture does not support any concrete OS for SDN based IoTs and the solution for wireless infrastructure based network does not fit in the infrastructure-less plethora of WSN. A summary of SDN based solutions for sensors networks is presented in Table 2.

D. Software defined Radio

The management of lower layer of the protocol stack is already introduced as Software Defined Radio (SDR) for managing the underlying complexity of hardwired implementation of the wireless network. The constituent entity of wireless communication is radio frequencies. With the increasing complexity and aggregated telecommunications technology and Radio Access Network (RAN) cross functionality is hard to obtain the desired result and need to physically intervene in radio technologies. By providing software-based radio manipulation, distinct management flexibility can uplift network performance. Constantly increasing IoT devices in billion and trillion and their communication need hardware independent implementation of network and radio connectivity.

SoftRAN [33] is proposed by Tomovic et al. which uses SDN principle in 4G LTE network. A centralised control plane abstracts the whole RAN into the geographical area. This Geographical area acts as a big base station where many radio elements i.e. physical base station are deployed under the control of the centralising controller; who manage radio resource allocation in the big base station.

The author proposed resource allocation in the form grid of three dimensions i.e. space, time, and frequency slots. The interaction between controller and radio element is done through APIs. Radioelement backup the information in the control plane. Based on this information, the controller decides to allocate resource in the domain of frequency, time and space slot. Radioelement takes some of its decision based on local information to manage the delay between controller and radioelement. Hence global network decisions are taken by controller local small resource management is done by the radio element.

SoftCell [34] incorporate SDN in the cellular core network and provide fine-grained policies for an LTE network. The contributing components in SoftCell architecture are i). Controller, ii). Access switches, iii). Core switches and iv). Middle-boxes. The controller defines policies and implement through switch level rules through middle-boxes. Traffic classification is done on the access switches. Every access switch has a local agent which caches each UE profile. In this way, local agent control of packet classification is access switch and undue burden over the controller is reduced. Controller has a global view and defined rules on the match fields i.e. policy tag, hierarchical IP address and UE identifiers.

The location and policies are embedded into packet header to avoid reclassification of the traffic. Core switches connect to the Internet through gateways fine-grained policies ensure through multi-dimensional aggregation and packet classification in asymmetric topology.

An integration of SDN and SDR in 5G network is proposed in [35] called Hybrid SDN/SDR architecture . The proposal architecture is cross layer combination of SDN and SDR for exploiting frequency spectrum and link information in 5G network. Network environment consists of spectrum and bandwidth perception in SDR layer while SDN controller can detect channel usage in the network. The cross-layer controller has used request frequency spread spectrum and is the decision maker and review flow traffic. This architecture also manages user authorization in the cross layer controller and grant access to a better band. The process of cross-layer communication between SDR and SDN starts with scanning spectrum holes.

SoftAir [36], proposed by Akyildiz et al. for the integration of SDN principals in 5G network by exploiting cloudification and network virtualization of a resilient network. The architecture provides mobility aware load balancing and resources efficient allocation through virtualization. The network architecture is based on software-defined switches and BSs which be dynamically programmed. The aggregated control is provided by NFV creating multiple virtual networks with independent protocols and resource allocation algorithms. Data plane comprises of SD-RAN and SD-core network nodes, which are OpenFlow-enabled. Data plane monitoring is done through OpenFlow and Common Public Radio Interface (CPRI). All management policies are defined at central control plane, which enables cloud orchestration. Traffic management module in control plane selects an optimal path in mobility aware context. QoS applications are carried out through distributed traffic classification module in the control plane. Overall, SoftAir presents a detailed and complete architecture of 5G cellular network management based on SDN and provide end-to-end QoS guaranty.

SDN&R [37] present a merger of SDN and SDR for IoT network and provide integrated management of diverse IoT network. SDN decouple the control plane from data plane and SDR is used to maintain radio status information in the control plane implemented on a base station (BS). The OpenFlow-enabled control plane performs radio control on the BS and cognitive edges (CE). The CE obtains the complete view of the radio spectrum. The packet processing is done on the controller connected to BS via a secure channel. The SDN-enabled cognitive radios resource management. This architecture is the detailed footprint of SDN integration in a cellular network for managing resources that are highly demanded in IoT network. A comparative review of studies literature Cellular IoT Solutions on SDN basis are presented in Table 3.

A. SDN based IoT Management

In a heterogeneous network like in IoT, where diverse technologies are interplaying and exchange information. In such networks, the management becomes very complex. The configuration, reconfiguration, resource allocation and even the pattern of intercommunication becomes extremely difficult. SDN, due to its decoupled nature, separate control plane from data plane offer programmability and management from a centralised server having a global view of the network status. SDN play a vital role in the management of such heterogeneous network. M2M communicating devices are managed through leveraging SDN control plane in [28]. The proposed framework is a two-tier architecture consisting of control plane and data plane and devices are IP enabled. These devices are populated with routing table as in the SDN-enabled switches. Controller has a complete view of the network. If a breakdown observed between devices and gateways, the controller does network reconfiguration. The communication between devices is used three reference points Mx, Gx, Gnx. The device kept its information and its neighbour information in the form of a file such that any change in the file is manipulated on controller instruction.

The management of a heterogeneous smart environment is quite complicated compared to a homogeneous M2M communication. Boussard et al. [53] proposed SDN based control and management framework for IoT devices in a smart environment. In their management framework, called "Software-Defined LANs (SD-LAN)", devices are organised and grouped in the order of requesting services from the user. The framework is a four-layer architecture consisting of (i) task description (ii). Service description, and (iii). Flow scheduling and low-level communication. This framework uses Universal Plug and Play (UPnP) and Simple Service Discovery Protocol (SSDP) discovery for the incoming device in the SD-LAN network. A virtual topology is created for SD-LAN devices based on services requirement such as audio, video, online game streaming etc.

The legacy routing waste resources and uses link unfairly. In the case of packet loss, the correlated latency also increases with caused performance degradation. In wired network packet drop may be caused by congestion on the link but in large-scale IoT devices (mostly wireless), this re-routing cause a Ping-Pong situation and the overall network performance degraded in case of any packet drop detected whether caused by a small interval. Context-aware IoT architecture

IoT applications occupy every domain of life and effect socio-economic factors such as health care, security, disaster management, remote access to things etc. In this context, D2D communication and coordination can play an important role where devices can seamlessly configure and reconfigure network without human intervention. Environment monitoring can be done if the IoT objects are implemented in a context-aware mode of communication. In [15], G. Savarese proposed a context-aware framework for LTE communication for D2D.

TABLE I. THE COMPARISON OF EXISTING SDN BASED MANAGEMENT SOLUTIONS FOR IOT

Architecture	management	Architecture	Control/data plane decoupling	Protocol used	scalability	Simulation Tools	benefit	Limitation
MINA[21]	Flow scheduling and management	Redefining the controller architecture based on DDS middleware and decouple the services and actual mechanism of traffic forwarding	Centralized controller	OpenFlow like protocol and IP protocol	-	Qualnet	Better performance and flow scheduling	layered controller design is critical to the management and still not addressed
Publish/subscriber-SDN[19]	Services/application management and resource management	It uses modular approach and translate user services message into SDN flow using DDS at the gateway	Centralised controller on the access point	COAP and OpenFlow	High	-	Scalability, mobility and security. Efficient handover	No validation proved through experiment or simulation results
CASSOWARY[20]	Profile and policy management	Context-aware sensor deployment using cassowary middle box on SDN controller. Network is divided into the In-Memory data grid	Device controller smart equipment	AMQP	Medium	cloudSim/ cassowary written in JAVA	Energy efficient and security profile and authentic access	Scalability

TABLE II. THE COMPARISON OF EXISTING WSN- SDN SOLUTIONS

Architecture	management	Architecture	Control/data plane decoupling	Protocol used	scalability	Simulation tools	benefit	Limitation
SDN_WSN[26]	Topology discovery and management	Centralised controller with three reference points	M2M communication between centralised controller and node	OpenFlow	Low	-	Intercommunication between devices and sensor node using gateways and centralised controller	Undefined functionality and implementation, no proof of evaluation.
WSN-SDN[27]	Sensor network flow management	WSN cluster with centralised controller monitored and controlled by Master SDN controller	Centralised master controller	OpenFlow/ distance aware routing protocol	Low	MATLAB	Optimal path selection, routing strategy adjustment on the network condition	Implementation of master and central controller is not clear, No proof of validation,
SD-WSN[28]	Infrastructure management and reconfiguration of sensor network	FPGA	Micro-controller	COAP	Low	-	Programmable reconfiguration of network	Hardware bounded and device dependency
ECCKN [29]	Energy management	Dumb data plane node dynamically associate with centralised controller where energy efficient algorithm ECCKN run to calculate routing on the basis of residual energy	Centralised controller with dumb data plane	ECCKN and OpenFlow	Undefined	-	Reduced total transmission time and centralised control	SDN implementation is not clear and protocol interaction is not specified
Senshare [30]	Open access Infrastructure management	Decoupling between infrastructure and	Dedicated overlay controller	Collection tree protocol (CTP)	Low	-	Support for multiple sensing	SDN controller implementation is not clear on overlay network

		application					applications reduced cost	
Integrate WSDN[31]	Management platform for using virtual machine in-network Processing (INNP)	Local controller in each sensor node which interacts with a centralised controller. INNP is done through VM in the node platform	Centralised controller and local controller	Contiki OS on each local controller	Low	Packet tracer trace the footprint of messaging and LQE	Flexible using commodity off the shelf device, reducing cost	Missing evaluation for behaviour and performance of WSN
SOF [32]	Flow management	INNP in data plane and flow-based packet forwarding	Centralised controller and distributed data plane	Sensor OpenFlow (SOF)	Low	-	handling peer compatibility, address classification, reduce setup latency, high throughput	Theoretical idea and not experimentally proved
SDN-WISE[24]	Localisation of distributed sensor in a centralised controller, energy management,	Centralised controller with dumb sensor node having flow table like OpenFlow flow table which is preinstalled with flow rules	Centralised controller, dumb data plane	OpenFlow	medium	-	The state-full approach, reduce information exchange. Mobility, reconfiguration and localisation of	Lacking security and reliability. In-depth architectural details are missing

This paper briefly describes the LTE network, M2M communication and an integration of LTE and D2D based on SDN in term of context-aware monitoring of LTE eNodeB that is responsible for allocating radio resources and scheduling traffic according to the QoS LTE network.The collected contextual information of LTE network, in the case of link failure or change in the network, is sent to the Context Aware Application (CAA) running on the M2M server where SDN controller can react to this change, reconfigure LTE network and allocate LTE resources in a flexible fashion.

Jararweh et al. in [10] proposed a comprehensive framework model for software defined system for IoT for the management and control of IoT devices in the heterogeneous network. The main focus is on the storage and security issues created in heterogeneous IoT network. The data generated and collected in IoT environment is immensely high which create storage issues. Some solutions propose the use of virtualized/software storage like in [38], where physical storage is abstracted by software storage and build the storage control operation in the centralised controller. Jararweh et al. use this architecture into the IoT environment. The main idea of collecting data from the sensor board which is aggregated on the IoT Bridge and send to SDSec controller for security checking. They use authentication and authorization for ensuring only authorise access. Afterwards, data is sent to IoT controller for rules definition for the collected data with the help of routing and controlling policies from SDN controller. And these rules are stored in the SDStore module of the framework which is used by the different application.

Much of the work has been done for the migration of IoT from a legacy network to SDN. In this regard, much-cited paper [21] by Qin et al. who proposed IoT architecture for flow scheduling based on Multi-network Information Architecture (MINA) with layer SDN controller. (The IoT tasks are usually depicted in an abstract manner and they are independent of underlying network and device resource specifications). In this proposed architecture, the authors proposed semantic modelling for the high-level task and low-level resource specifications and represent IoT task as hierarchal semantic task and parameters are written in term of ontological concepts. The Task plans are stored in task Knowledge Base and resources with capabilities are stored in resource Knowledge Base. The IoT task is matched with task KB and submits to an analyser, which extracts both KBs, find resources with capabilities, and provide and appropriate solution, which is then mapped with the service solution specification. Information for resource mapping is obtained from Network information Base or DB. Afterwards, flow scheduling is done on the basis of state information provided by MINA state global information view. The QoS service is analysed using network calculus model and path is obtained by using Genetic Algorithm (GA) where each flow has a chromosome, which is a path between source and destination, and genes are considered as nodes on that network. The implantation is done in the Qualnet simulator by taking smart campus network topology. The performance metric used to delay, throughput and jitter for file-sharing, tele-audio, and video flow over the network and compared their GA scheduling with two existing SDN scheduling algorithms bin-packing and load balance algorithms and find that their results are consistent. In this paper, the author et.al did not found the flow entry overhead in the beginning and consider that their flow scheduling GA is stable.

However, the initial overhead is not negligible and it is assumed that the flow is proactively registered in the controller. In the case of wireless IoT device, there is a chance of change in the topology, which needs to reregister the flow, which create extra overhead and performance degrade.

In [55], Xiong et al. presented resource allocation architecture for SDN based IoT network. The average reward of the network is increased by considering long-term expected average reward per unit time and based on this reward optimal

resource allocation problem using MDP. The reward model is computed by assuming states and actions in each state. Using this reward, an optimal resource allocation policy is formulated using value iteration algorithm.

Ancuta. et al in [56] presented the concept of a management solution for dynamically instantiated services in an elastic environment. The information is exchanged between different entities consuming more energy when HTTP protocol is used for message forwarding. In this context, an extendable architecture open MTC is proposed and its implementation is prototyped which uses oneM2M and ETSIM2M protocol that run on Gvent API. They show that as soon as the new instance in M2M arrives, the information is an exchange between M2M management adaptors which informed the M2M connectivity manager who retains the policies for the M2M devices. This transport policy is announced. By this implementation, the scalability can be increased but there is a factor of delay as the number of devices increased in the network.

B. SDN-Based IoT Operating System/controllers

The IoT devices, in general, are heterogeneous and use multiple technologies for intercommunication. Even though IoT uses multiple middlewares to reduce, the gap between application and IoT devices message passing, interoperability is still an issue to enhance the performance and increase the reusability of IoT network. To deal with this interoperability, network Operation System (NOS) play an important role in managing interoperability in heterogeneous systems. As sensor nodes and actuator are considered as a building block of an IoT network. These tiny device/motes are constraints of energy resources, storage capacity, and processing power, content-based routing etc.

However, the established OS for these tiny IoT components in a WSN based IoT network are not capable of handling interoperability on large scale and conversion of flow. For this reason, many OS, Such as Contiki [57], RIOT OS [58], Tiny OS [59], Lite OS [60] etc. were presented for WSN based IoT network. However, these operating systems are specific to the certain application, thus lacking flexibility and dynamism i.e. independent of platform in a system. A comparative analysis of these all OS is presented in Table. 4.

Still, there is no concrete OS for managing the integration of IoT and SDN. In this context, a little effort is put in developing OS for SDN based IoTs which in return create complexity in translating flow rules/policies for IoT devices. SDN is also in its infancy and it uses OpenFlow is used for bridging gap between SDN control plane and data plane. Few NOS are also available in market such as NOX, ONIX [62], Maestro [63], OpenDaylight [14] etc. These controllers are well operated for wired SDN but these OS are not suitable for SDN-driven IoT network. This controller or OS lack support for the characteristics of IoT devices such as fundamental energy and processing constraints, data aggregation, duty cycle etc. The initiating concept of reprogramming and re-tasking in WSN was proposed in Sensor OpenFlow (SOF) [47]. SOF is three layer architecture; application layer, a control layer and data plane layer. The application layer consists of all applications necessary for managing query applications, data processing applications etc. Control layer consisting modules are "sensor re-configuration" module and "query strategy control" module and perform flow-based forwarding in the data plane consisting FDs sensor nodes. Forwarding plane forwards the sensor flow in the order defined by the controller. However, flow creation and management was a challenge in SOF and the overhead created due to control traffic can dim the expected outcome of SOF. To overcome these limitations, complexity is added and simplicity is reduced. For the sack of providing flexibility and simplicity in WSNs through SDN, an operating system solution based on SDN was proposed by Galluccio, et al. in [38], named as SDN-WISE; an architecture and operating system for WSN support duty cycle and data aggregation and provide a state-full solution for SDN. The consisting data structures of SDN-WISE are the WISE States Array, the Accepted IDs Array, and the WISE Flow Table. The communication between sensor nodes and other controller is done through WISE-Visor resemble in the functionality of FlowVisor [65] which is switching virtualization approach in SDN. The introduced adoption layer performs translation between the sensor node and WISE-Visor and decouples data plane and control in the SDN based sensor network. SDN-WISE is a state-full approach and defines its policies on the basis of state description, shown in adopted example from [66] which depict policy implementation for a packet if its threshold or measure is less than a certain threshold (X_{thr}) and it is generated by node A as shown in Fig. 6. Details of the studied literature in the context of the controller and operating systems in sensor networks are given below in Table 4.

SDN-WISE ensure the minimum number of information exchange and holistic support for different protocols and node design. Christos et al. do an enhancement in SDN-WISE in [67]. The authors propose an OS based on Open Network Operating System (ONOS) [68] and integration of SDN-WISE and OpenFlow network in a seamless manner. An OpenFlow-enabled device can interact with a WSN network through ONOS.

C. SDN security framework for IoT

In the most recent IoT arena, billions of Internet-connected physical objects produces the bulk of data within few milliseconds whose storage, processing, automation, and management is an intensive task. These devices are potentially under threat due to unbounded connectivity and communication over wired and wireless transmission medium due to lack of standard security protocol/architecture for IoTs. SDN is considered a powerful technology of having centralised control over the information flow in the network and provide a preemptive security policy. The IoT system becomes more vulnerable to security risks when they are monitored from a centralised controller as SDN based IoT network.

Little considerations of security aspect are witnessed in SDN based IoT network. In [39], Sahoo et al. proposed a secure architecture for IoT network based on SDN. There are five basic security properties which need to be under consideration while defining a security model. These security characteristics are Confidentiality, integrity, availability, authentication and non-repudiation [39]. Sahoo et al. proposed

their secure architecture on the basis of authentication of IoT device on the controller. In this architecture, the considered IoT is an ad hoc network in which wireless object establish a connection with the controller and controller block all the port when the connection is established and controller starts authentication. If the user is authentic, the controller starts pushing flow to that user. Few controllers in the network serve as a security guard and exchange information with each other about the user authentication. In the case of guard controller failure, some other border controller is selected as security controller.

Even though this work presents a basic layout for secure SDN based IoT network, however, the validity and correct operation are not provided.

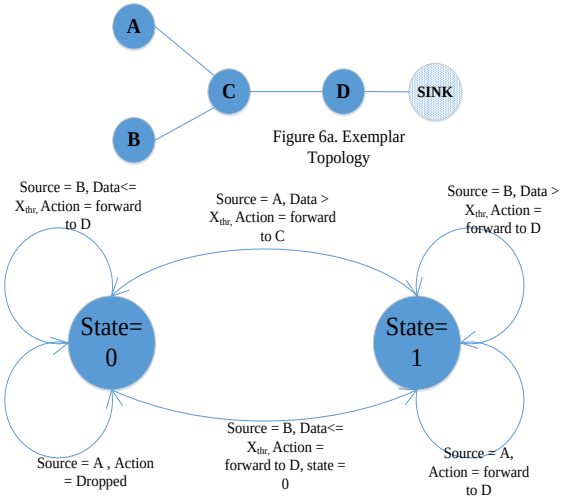


Figure 6a. Exemplar Topology

Fig. 6. Action derived from FSM in SDN-Wise

In [40], authors proposed a dynamic firewall named as Distributed Smart Firewall (DISFIRE) for secure architecture in SDN based grid network. The architecture consists of hierarchal cluster network with multiple SDN controllers. These cluster head SDN controller implement a security policy. For this purpose, they used cisco defined policy agent opFlex [41] in the controller instead of OpenFlow. The device information is exchanged between devices and any unauthorised potentially malicious device flow rule policy is deleted.

A security proposal for smart cities is presented in [42]. Chakrabarty *et al*. proposed a secure architecture based on trusted SDN controller, Black Network, Unified Registry and Key Management System in an IoT network. The security architecture ensures authentication of the heterogeneous devices. SDN controllers act as a Trusted Third Party (TTP) and provide security properties i.e. confidentiality, privacy, integrity, authentication, and routing between IoT devices. The unified registry is responsible for Identity management, availability, accounting, authentication, authorization. The shared key is used for secure communication.

Ad-hoc network in term IoT network do not provide access control and traffic monitoring in ad hoc network is not possible therefore security is a threat in ad hoc network where infrastructure is missing and connection are established reactively. In [43] Architecture is presented where each node is connected to a domain controller through an embedded virtual switch. This controller is on the edge of the network and acts as a domain controller and provide authentication of the network devices. On the authorization profile, flow entries are pushed in the access switch. Oliver *et al*. [44] proposed a SDN based IoT architecture for infrastructure and infrastructure-less network where a virtual switch is embedded in each node bounded to a controller in a domain. Devices in different domains interact with the border switch. Some of the border switches are selected as controller and these controller acts as a security controller. The security controllers provide dynamic network configuration and security policy deployment. The architecture provides Authentication of the network devices on the time of device registering with the controller.

IoT/M2M communication can leverage emergency response in case of network failure in a disaster situation and can aid the first responder in taking appropriate decisions. In [45], a security architecture for the first responder in the IoE/IoT environment is proposed using Software Defined Perimeter (SDP) protocol. Where SDP collect the IP addresses of all M2M communication capable devices and store into a logical network. When any new M2M device comes in close proximity of SDP domain, they first configure themselves in a secure SDP by using authentication credentials. SDP efficiency of authenticating secure access in the emergency response is visible, it also can data privacy and trust in the M2M communication network.

SDIoT [10] present the security of SDN based IoT network by implementing SDSec module which utilised NFV to create a virtual topology for the connected device and leverage the benefit of SDP for authentication by block all the switch port when received a request from a new flow. SDSec store information in the security database and it identifies an object by tracking authentication DB. SDN controller set flag P if everything is good otherwise flag N for negative. If the flag is set P then flow is allowed to enter and access is granted. Another security framework is proposed in [46]. In this architecture, author uses IoT agent and IoT controller that are responsible for connecting SDN controller in the SDN-enabled heterogeneous network. IoT agent is registered agent with IoT controller. SDN controller performs authentication and routing based on collected information from the IoT agents. The whole IoT network is divided into segments with its own SDN controller. Every IoT device must be connected to an OpenFlow enabled IoT device, which coordinates with segment controller. The inter-segment communication is through gateway controller. Embedded system implication in intensive health monitoring is a rich field; highly requiring security and reliability in information interchange. Cyber-attacks and malicious encroachment are very common in the Internet-connected environment and can modify the functioning of embedded systems. Security system in the embedded system does not entail high processing security techniques. Ukil *et al*. exploited the detail security threats in embedded system in [47]; proposed Secure Execution environment (SEE) mediating security model from outside security threats. Dedicated security processor

compartmentalised from non-secure mode is SEE architecture with dedicated RAM for retaining integrity and confidentiality from out the SEE code. Intrusion detection system (IDS) implementation is not easy in IoT as it requires complex mathematical computation and profile based modelling. In [48], Skowyra *et al.* exploited the idea of IDS based learning in the mobile embedded system for restraining modification from any anomaly either from inside the network or from out of the network. The OpenFlow controller contains all logic and defines rules based on state-full information. Table. 5 presents the studies literature about the security-related solution in IoT-based on SDN.

IV. Discussion and Open Issues

The whole concept of IoT-SDN is not mature, and standardisation efforts are still under way, multiple competing alliances are trying to dominate for a global standard. We have discussed broad literature on the integration of SDN and IoT. In this study, different aspects of SDN integration in IoT technology in the context of M2M communication, LTE/IoT communication, Sensor IoT heterogeneous network are discussed. It also highlights the proposed solutions for architecture, management framework; security aspect in the SDN based IoT. A detailed overview of the observed studies is given in Table 5; which demonstrate the diversity of SDN incorporation in different IoT domains. Another important thing to notice that most of these studies are not experimentally validated; however only a representative proposal frameworks are grabbing the attention during last five years. This is because of the anticipated benefits of SDN programmability in the management of mushroom growing IoT devices. This effort could become a reference point for the researchers and developers to investigate the trending IoT application in a more controlled way; proving fast innovation and change due to technology shifts.

However, the existing solution is not fully integrated into SDN and a comprehensive architecture and framework are not established so far. Few effort are really admirable such as SDIoT, BlackSDN etc.,

where a complete framework for IoT devices is presented giving SDStorage, SDSystem and SDSec for management, security and architectural detail of IoT interplay in SDN. A major factor of lacking a comprehensive architecture for SDN based IoT is the absence of a concrete framework of IoT architecture.

SDN main characteristics lie in the wired and infrastructure-based network, while in IoT, devices are diverse in nature and different communication technologies are blended to form a heterogeneous network. This merger may be mobile in case of ad hoc network or vehicular network where dynamic allocation of resources with constraints devices need object addressing, which is still not addressed in SDN, based IoTs

Another issue in the IoT network is content addressing and context awareness in services provisioning with QoS support, which is still not addressed in any work. Existing transport protocols fail in the IoT scenarios since their connection setup and congestion control mechanisms may be useless; furthermore, they require excessive buffering to be implemented in objects. Also, the traffic pattern in IoT is different from the traditional network traffic even different from SDN/OpenFlow data flow; require excessive intention from the researchers.

In SDN, loT of control traffic consume bandwidth and hence degrade the spectral efficiency in the IoT Devices. Also, the battery power is highly vulnerable to this massive control traffic. In IoT devices, the traditional security characteristic is hard to implement, the authentication and authorization require a storing of authentication profiles in the minute storage. Well-known traditional network security cannot be applied in IoT. SDN centralised control plane may suffer from denial of services attack and man in middle attack. Due to the huge amount of data produced in IoT network, data privacy is a critical issue in the case of M2M communication in IoT network.

The controller is still not defined for the IoT. The controller took a lot of space and implemented on the server side; in that case, the instruction set produced by the SDN controller should be formatted according to the IoT devices. The single centralised controller is prone to single point failure; therefore, a need for distributed controller is a research question in IoT communication network.

V. Qualitative Predictions for 2020

The IoT will help in establishing smart ecosystems such as smart home, smart building, smart health care unit, disaster management, smart industrialisation, nifty transportation and smart grid station etc. and eventually bring a social and industrial revolution. According to a statistic data obtained from [2], around 14.4 billion connected devices were there in 2014 and will reach up to 50 billion connected devices in 2020. The increasing trend in the IoT connected device with respect to the world population is shown in Fig. 7

IoT adopting is like a wildfire spreading across dry grass and millions of IoT-enabled smart devices are in operation like sensors, actuators, RFIDs, vehicles, PDAs, smartphone, cellular devices, wearable's, smart bulbs, smart turbines, smart arms and much more. This widespread adoption of smart object and interconnectivity has changed the market and research interest. According to a report by Gartner, Inc., around 6.4 billion devices are in play till 2016 which is 30% more than in 2015 and there is approximately 5.5 million new devices are connecting to the Internet per day. This count is immense increased and will reach to around 20.8 billion in 2020 (according to Gartner report) and will reach up to 50 billion connected devices in 2020 creating revenue of $14.4 trillion. Due to this high-expected statics, companies are bullishly spending a huge amount on IoT integration; around $656 billion were spending in 2014, which estimate a rise up to $1.7 trillion in 2020. It is estimated that there will be a 90% rise in the installation of intelligence and smart connectivity in cars until 2020, which was only 2% in 2012. This swift switch is forcing manufacturers and industries to look into broader sense and hence research trends are changes as shown in Fig. 8.According to International data corporation, around $8 billion will be generated which was only $960 million dollars in 2014; 90% compound growth rate. According to Gartner,

SDN application and infrastructure is top 10 strategic during 2015. The annual data growth rate also crosses limits in zeta-bytes in 2016 and predicted to cross up to 2.3 ZB by 2020. According to IDC, overall enterprise network revenue will grow 3.5% to reach $41.1 billion When the growth rate comes in term of SDN then according to Gartner report there is 87% increase in production in the data centre using SDN and revenue generated was $960Million in 2014and will raise to $8Billion by 2018 i.e. 734% a total rise. The increase in both domains clearly predicts a merger of two technologies and increase in the SDN based IoT production.

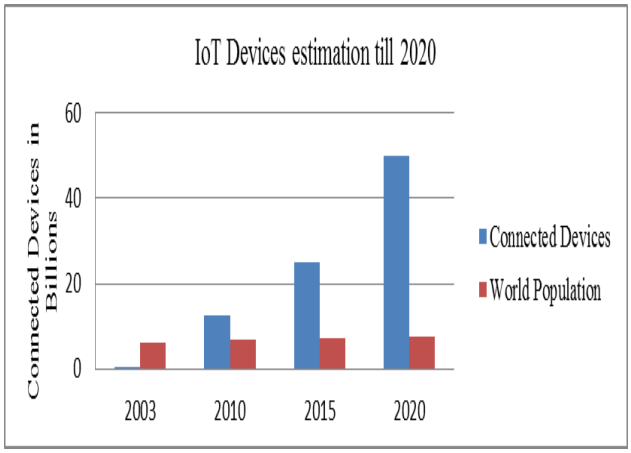


Fig. 7. Worldwide IoT connected devices

TABLE III. THE COMPARISON OF EXISTING CELLULAR IOT SDN SOLUTIONS

Architecture	Target network	Resource management	Interface API	Purpose	Cloudification/virtualization	Control/data plane decoupling	Traffic engineering	Scalability	Simulation tools	Benefit	Limitation
SDR	Wi-Fi ,WIMAX	Spectrum management at software level	Smart antenna API	Increasing the spectrum allocation efficiency and virtual network/slices to support multiple wireless protocol instances	No virtualization	Centralized controller	uplink and downlink spectral efficiency	Low	MATLAB, Simulink	Independent of physical spectrum allocation	Unpredicted user behaviour, inflexible traffic engineering, Advanced spectrum management
SoftRAN[33]	5G/ LTE	Resource management, mobility support, traffic offloading	Controller API/Femto API	To overcome the tightly bounded co-ordination in resource management	Big base station	Centralised controller and local agent at eNBs , Abstraction through slicing forming big base station	Load balancing Interference management/ SD Radio access network	Low	LTE-SIM	Radio resource management , mobility support, Traffic offloading, Reduced delay	No concrete solution, virtualization is not clear, centralised control plane and interaction between core network and RAN is not defined
Hybrid SDN-SDR[35]	5G	Spectrum management	-	For the management of spectrum allocation and network management (e.g., bandwidth) in a 5G network	No virtualization	Centralized controller	Spectrum resource management and network resource management	low	MatLab	Power saving and optimisation	Cross-layer controller, security
SoftCell [34]	Cellular network	Fine grain policies management.	Open Flow API	Modification in the core network	Minimum virtualization	Logically centralised controller, local	MPLS and slandered routing as in OpenFlow	high	SoftCell implemented on Floodlight	Dynamic traffic offloading, efficient routing,	Fine grain service policies

						agent SD-RAN (BS)			controller and micro-benchmarking using bench	minimising the state in the core network	
SoftAir[36]	5G	Distributed traffic classification, fine grain virtualization, network management (routing)	Open Flow & CPRI	network function cloudification and network virtualization	Fine grain virtualization	SD-BS, SD-switch, BS-clustering	Collaborative processing, scheduling and mobility management	high	-	Flexible platform for fully & partially centralised architecture	Security issue not addressed
cellSDN [16]	Cellular network	Mobility management and policy control management	NOS	virtualization	Basic support for virtualization	Centralised control plane, local control agent ate BS	MPLS traffic labelling or VLAN tags	Low	-	Seamless mobility management and fine grain control due Local agent	No proof of concept and evaluation of the proposed scheme, vague traffic engineering handling using MPLS/VLAN tags

TABLE IV. WSN BASED OS IN IOTS

Operating System	Action	Programming Language	RAM required (Kb)	Kernel Implementation	Service management	Kernel management	Model
Contiki	Event based	C	2	Preemptive multithreading	Dynamic	Hybrid	-
RIOT OS	Task based	C/C++	1.5	Multithreading	Dynamic	Static	-
TinyOS	Event	NesC	1	Partial	Static	Dynamic	Concurrency model
Lite OS	Event based	C	4	Multithreading	Dynamic	Dynamic	hierarchical file system
SDN-WISE	Event based	Java	10	State-full	State-full	Dynamic	Modular
ONOS	Event based	Java	8	-	-	-	modular

TABLE V. SDN-BASED IOT SECURITY SOLUTIONS

Approach	Security parameter	Network	description	Limitations
secured SDN framework **[39]**	Authentication	Ad hoc network	SDN controller block all switch port on receiving new flow and start authentication	Not prove implementation or simulation, only a theoretical framework
DISFIRE**[40]**	Authentication & authorization	Grid network	hierarchal cluster network with multiple SDN controllers implement a dynamic firewall to ensure authorization	Evaluation of framework lacking. The protocol used is opflex which is not practically tested
Black SDN**[42]**	Location Security, Confidentiality, Integrity, Authentication And Privacy.	Generic IoT/M2M communication	secure the meta-data and the payload by encryption in the link layer and use SDN controller as TTP	Scalability in black network will create hazard in providing complete security
SDP**[45]**	Authentication	Ad hoc network/M2M communication	SDP collect the IP addresses of all M2M communication capable devices and store into a logical network. And authenticate on the basis of information stored	Scalability will encounter performance in case of IoE
SDIoT**[10]**	Authentication	Generic IoT network	It utilised SDSecurity mechanism leveraging NFV and SDP for ensuring secure access in the network by authentication.	Hard to manage the large network in case of single SDSec logical element. An experimental evaluation is lacking
[43][44][46]	Authentication, security policy at security controller	Generic IoT	Domain controller and edge controller for security and intercommunication between	Lacking proof for concept, not tested not evaluated

			different domain/segments	
SEE [47]	Confidentiality, Integrity	Embedded devices/System	Theoretical concept of encountered security threats in an embedded system	Processing slows down
L-IDS [48]	Learning network IDS	Mobile embedded devices (MEB) for the institutional site.	Mobile embedded devices dynamically form connection with the infrastructure where the possible attacker can attack MEB and	A Large number of control message interchange creates congestion on the controller. Experimental validation in not done yet.

TABLE VI. DESCRIPTIVE SUMMARY OF IMPORTANT SDN-IOT SOLUTION FRAMEWORKS

Approach	purpose	Implementation domain	Year	Operating system/controller
SDN-6LoWPAN [49]	NFV for bandwidth utilization	IPv6 local WPAN	2015	Centralized SDN controller
SDN-M2M [50]	Network configuration and resource management	M2M communication devices	2014	Centralised SDN controller
MINA[21]	Flow scheduling and management	Middleware	2014	Centralized controller
Publish/subscribe-SDN[19]	Services/application management and resource management	Generic IoT	2015	Centralized controller
CASSOWARY[33]	Profile and policy management	WSN	2015	Centralized
SDN_WSN[46]	Centralized controller with three reference points	WSN	2014	Centralized controller
WSN-SDN[41]	Sensor network flow management	WSN	2014	Hierarchal controller (cluster and master controllers)
SD-WSN[42]	Infrastructure management and reconfiguration of sensor network	WSN	2014	FPGA microcontroller
ECCKN [29]	Energy management in sensor network	WSN	2016	Centralised controller with dumb data plane
Senshare [44]	Open access Infrastructure management	Sensor networks	2012	Dedicated overlay controller
Integrated WSDN-[45]	Management platform for using virtual machine in-network Processing (INNP)	WSN	2015	Local and centralised controller
SOF [47]	Flow management	WSN	2012	Centralised controller and distributed data plane
SDN-WISE[38]	Localisation of distributed sensor in a centralised controller, energy management	WSN	2015	Centralized controller
SDR	Spectrum management at software level	Wi-Fi ,WIMAX	2012	Centralised control plane
CellSDN[16]		Cellular network	2012	
SoftRAN[33]	Resource management, mobility support, traffic offloading	5G/LTE	2013	Big base station
SoftCell[49]	Fine grain policies management.	Cellular network	2013	Logical centralized controller
Hybrid SDN-SDR[35]	Spectrum management	5G	2014	Centralized controller
SoftAir[36]	network function cloudification and network virtualization	5G	2015	SD-Centralized controller
secured SDN framework [39]	Authentication	Ad-hoc networks	2015	SDN controller block
SDP[45]	Authentication	Ad hoc network/M2M communication	2015	Central controller and local agents
DISFIRE[40]	Authentication & authorization	Smart Grid network	2016	hierarchal cluster network with multiple SDN controllers
Black SDN[42]	Location Security, Confidentiality, Integrity, Authentication And Privacy.	Generic IoT/M2M communication	2016	Centralized controller
SDIoT[10]	Authentication & authorization	Generic IoT	2015	SDSec module on SDN controller
SEE [47]	Confidentiality, Integrity	Embedded system	2011	-
L-IDS [48]	Learning network IDS	Embedded system	2013	OpenFlow controller

VI. CONCLUSION

IoT is a new norm of connectivity, enabling smart ecosystem. It is changing the way we think to communicate with an object in out surroundings and improving the quality of life. However, IoT lacks programmability, agility, security and data management due to the huge amount of data produced. To meet the need of customer requirement, it is highly anticipated use programmability and centralised control for IoT management. In SDN, control plane and data plane are decoupled, which hide the high-level implementation of the low-level forwarding devices. In this paper, we have surveyed the existing solution for the integration of SDN control plane in IoT network. In this work, first, we have discussed the existing for the IoT management based on SDN centralised control plane in different IoT contributors, summarising architectural details and its evolution, and then outline the unresolved issues in this merger and reported some predictions for the world in 2020.

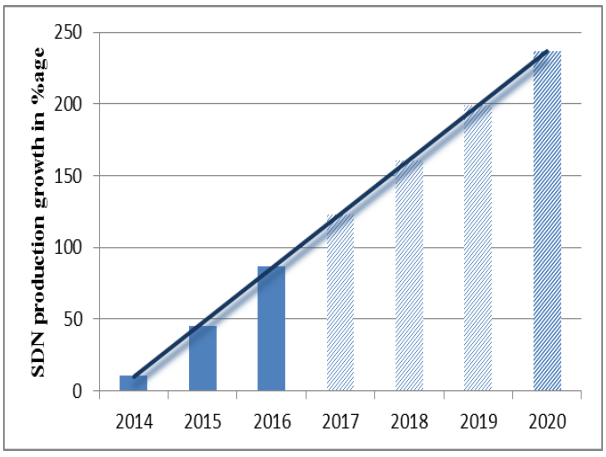


Fig. 8. SDN growth in data centers prediction for 2020

REFERENCES

[1] E. Borgia, "The Internet of Things vision: Key features, applications and open issues," Comput. Commun., vol. 54, pp. 1 – 31, 2014.

[2] D. Kreutz, F. M. Ramos, P. E. Verissimo, C. E. Rothenberg, S. Azodolmolky, and S. Uhlig, "Software-defined networking: A comprehensive survey," Proc. IEEE, vol. 103, no. 1, pp. 14–76, 2015.

[3] B. A. A. Nunes, M. Mendonca, X.-N. Nguyen, K. Obraczka, and T. Turletti, "A survey of software-defined networking: Past, present, and future of programmable networks," IEEE Commun. Surv. Tutorials, vol. 16, no. 3, pp. 1617–1634, 2014.

[4] Y. Jarraya, T. Madi, and M. Debbabi, "A survey and a layered taxonomy of software-defined networking," IEEE Commun. Surv. Tutorials, vol. 16, no. 4, pp. 1955–1980, 2014.

[5] A. Lara, A. Kolasani, and B. Ramamurthy, "Network innovation using OpenFlow: A survey," IEEE Commun. Surv. Tutorials, vol. 16, no. 1, pp. 493–512, 2014.

[6] A. Al-Fuqaha, M. Guizani, M. Mohammadi, M. Aledhari, and M. Ayyash, "Internet of things: A survey on enabling technologies, protocols, and applications," IEEE Commun. Surv. Tutorials, vol. 17, no. 4, pp. 2347–2376, 2015.

[7] L. D. Xu, W. He, and S. Li, "Internet of Things in Industries: A Survey," IEEE Trans. Ind. Informatics, vol. 10, no. 4, pp. 2233–2243, Nov. 2014.

[8] C. Perera, C. H. Liu, S. Jayawardena, and M. Chen, "A survey on Internet of things from the industrial market perspective," IEEE Access, vol. 2, pp. 1660–1679, 2014.

[9] Z. Yang, Y. Yue, Y. Yang, Y. Peng, X. Wang, and W. Liu, "Study and application on the architecture and key technologies for IOT," in Multimedia Technology (ICMT), 2011 International Conference on, 2011, pp. 747–751.

[10] N. Bizanis and F. Kuipers, "SDN and virtualization solutions for the Internet of Things: A survey," IEEE Access.

[11] T. D. N. Gray Ken, SDN: Software Defined Networks.

[12] N. Gude, ., Koponen, T., Pettit, J., Pfaff, B., Casado, M., McKeown, N., & Shenker, S., "NOX: Towards an Operating System for Networks," SIGCOMM Comput Commun Rev, vol. 38, no. 3, pp. 105–110, Jul. 2008.

[13] W. Braun and M. Menth, "Software-Defined Networking using OpenFlow: Protocols, applications and architectural design choices," Future Internet, vol. 6, no. 2, pp. 302–336, 2014.

[14] [14] J. Medved, R. Varga, A. Tkacik, and K. Gray, "Open daylight: Towards a model-driven sdn controller architecture," in Proceeding of IEEE International Symposium on a World of Wireless, Mobile and Multimedia Networks 2014, 2014.

[15] A. Shalimov, D. Zuikov, D. Zimarina, V. Pashkov, and R. Smeliansky, "Advanced study of SDN/OpenFlow controllers," in Proceedings of the 9th central & eastern European software engineering conference in Russia, 2013, p. 1.

[16] N. McKeown *et al*., "OpenFlow: enabling innovation in campus networks," ACM SIGCOMM Comput. Commun. Rev., vol. 38, no. 2, pp. 69–74, 2008.

[17] A. Doria, Salim, J. H., Haas, R., Khosravi, H., Wang, W., Dong, L., and Halpern, J. "Forwarding and control element separation (ForCES) protocol specification," 2010.

[18] H. Song, "Protocol-oblivious forwarding: Unleash the power of SDN through a future-proof forwarding plane," in Proceedings of the second ACM SIGCOMM workshop on Hot topics in software defined networking, 2013, pp. 127–132.

[19] Y. Jararweh, M. Al-Ayyoub, A. Darabseh, E. Benkhelifa, M. Vouk, and A. Rindos, "SDIoT: a software defined based Internet of things framework," J. Ambient Intell. Humaniz. Comput., vol. 6, no. 4, pp. 453–461, 2015.

[20] D. Evans, "The Internet of things," Evol. The Internet Is Chang. Everything Whitepaper Cisco Internet Bus. Solutions Group IBSG, vol. 1, pp. 1–12, 2011.

[21] "Internet of Things - Architecture — IOT-A: Internet of Things Architecture."

[22] "LTE Overview," www.tutorialspoint.com. [Online]. Available: https://www.tutorialspoint.com/lte/lte_overview.htm.

[23] J. Liu, S. Zhang, N. Kato, H. Ujikawa, and K. Suzuki, "Device-to-device communications for enhancing the quality of experience in software defined multi-tier LTE-A networks," IEEE Netw., vol. 29, no. 4, pp. 46–52, 2015.

[24] G. Savarese, M. Vaser, and M. Ruggieri, "A Software Defined Networking-based context-aware framework combining 4G cellular networks with M2M," in Wireless Personal Multimedia Communications (WPMC), 2013 16th International Symposium on, 2013, pp. 1–6.

[25] L. Erran, L. Z. Morley, and M. J. Rexford, "Cellsdn: software-defined cellular networks," 2012.

[26] M. H. Kabir, "A Novel Architecture for SDN-based Cellular Network," Int. J. Wirel. Mob. Networks, vol. 6, no. 6, p. 71, 2014.

[27] M. M. Mazhar, M. A. Jamil, A. Mazhar, A. Ellahi, M. S. Jamil, and T. Mahmood, "Conceptualization of Software Defined Network layers over Internet of things for future smart cities applications," on Wireless for Space and Extreme Environments (WiSEE), 2015 IEEE International Conference on, 2015, pp. 1–4.

[28] H. Huang, J. Zhu, and L. Zhang, "An SDN_based management framework for IoT devices," in Irish Signals Systems Conference 2014 and 2014 China-Ireland International Conference on Information and

Communications Technologies (ISSC 2014/CIICT 2014). 25th IET, 2014, pp. 175–179.

[29] G. Kim, J. Kim, and S. Lee, "An SDN based fully distributed NAT traversal scheme for IoT global connectivity," in Information and Communication Technology Convergence (ICTC), 2015 International Conference on, 2015, pp. 807–809.

[30] P. Bull, R. Austin, and M. Sharma, "Pre-emptive Flow Installation for Internet of Things Devices within Software Defined Networks," in Future Internet of Things and Cloud (FiCloud), 2015 3rd International Conference on, 2015, pp. 124–130.

[31] "Cisco Visual Networking Index: Global Mobile Data Traffic Forecast Update, 2015–2020 White Paper," Cisco.

[32] A. Hakiri, P. Berthou, A. Gokhale, and S. Abdellatif, "Publish/subscribe-enabled software defined networking for efficient and scalable IoT communications," IEEE Commun. Mag., vol. 53, no. 9, pp. 48–54, Sep. 2015.

[33] P. Kathiravelu, L. Sharifi, and L. Veiga, "Cassowary: Middleware Platform for Context-Aware Smart Buildings with Software-Defined Sensor Networks," in Proceedings of the 2Nd Workshop on Middleware for Context-Aware Applications in the IoT, New York, NY, USA, 2015, pp. 1–6.

[34] Z. Qin, G. Denker, C. Giannelli, P. Bellavista, and N. Venkatasubramanian, "A Software Defined Networking architecture for the Internet-of-Things," in 2014 IEEE Network Operations and Management Symposium (NOMS), 2014, pp. 1–9.

[35] D. Wu, D. I. Arkhipov, E. Asmare, Z. Qin, and J. A. McCann, "UbiFlow: Mobility management in urban-scale software defined IoT," in 2015 IEEE Conference on Computer Communications (INFOCOM), 2015, pp. 208–216.

[36] A. El-Mougy, M. Ibnkahla, and L. Hegazy, "Software-defined wireless network architectures for the Internet-of-Things," in Local Computer Networks Conference Workshops (LCN Workshops), 2015 IEEE 40th, 2015, pp. 804–811.

[37] M. Usman, A. A. Gebremariam, U. Raza, and F. Granelli, "A Software-Defined Device-to-Device Communication Architecture for Public Safety Applications in 5G Networks," IEEE Access, vol. 3, pp. 1649–1654, 2015.

[38] L. Galluccio, S. Milardo, G. Morabito, and S. Palazzo, "SDN-WISE: Design, prototyping and experimentation of a stateful SDN solution for WIreless Sensor networks," in 2015 IEEE Conference on Computer Communications (INFOCOM), 2015, pp. 513–521.

[39] A. D. Gante, M. Aslan, and A. Matrawy, "Smart wireless sensor network management based on software-defined networking," in Communications (QBSC), 2014 27th Biennial Symposium on, 2014, pp. 71–75.

[40] P. Dely, A. Kassler, and N. Bayer, "OpenFlow for wireless mesh networks," in Computer Communications and Networks (ICCCN), 2011 Proceedings of 20th International Conference on, 2011, pp. 1–6.

[41] Z. Han and W. Ren, "A novel Wireless Sensor Networks structure based on the SDN," Int. J. Distrib. Sens. Networks, vol. 2014, 2014.

[42] T. Miyazaki S. Yamaguchi, K. Kobayashi, J. Kitamichi, S. Guo, T. Tsukahara, and T. Hayashi,, "A software defined wireless sensor network," in Computing, Networking and Communications (ICNC), 2014 International Conference on, 2014, pp. 847–852.

[43] Y. Wang, H. Chen, X. Wu, and L. Shu, "An energy-efficient SDN based sleep scheduling algorithm for WSNs," J. Netw. Comput. Appl., vol. 59, pp. 39–45, 2016.

[44] I. Leontiadis, C. Efstratiou, C. Mascolo, and J. Crowcroft, "SenShare: transforming sensor networks into multi-application sensing infrastructures," in European Conference on Wireless Sensor Networks, 2012, pp. 65–81.

[45] M. Jacobsson and C. Orfanidis, "Using software-defined networking principles for wireless sensor networks," in 11th Swedish National Computer Networking Workshop (SNCNW), May 28-29, 2015, Karlstad, Sweden, 2015.

[46] H. Huang, J. Zhu, and L. Zhang, "An SDN_based management framework for IoT devices," in 25th IET Irish Signals Systems Conference 2014 and 2014 China-Ireland International Conference on Information and Communications Technologies (ISSC 2014/CIICT 2014), 2014, pp. 175–179.

[47] T. Luo, H.-P. Tan, and T. Q. Quek, "Sensor OpenFlow: Enabling software-defined wireless sensor networks," IEEE Commun. Lett., vol. 16, no. 11, pp. 1896–1899, 2012.

[48] A. Gudipati, D. Perry, L. E. Li, and S. Katti, "SoftRAN: Software defined radio access network," in Proceedings of the second ACM SIGCOMM workshop on Hot topics in software defined networking, 2013, pp. 25–30.

[49] X. Jin, L. E. Li, L. Vanbever, and J. Rexford, "SoftCell: Scalable and Flexible Cellular Core Network Architecture," in Proceedings of the Ninth ACM Conference on Emerging Networking Experiments and Technologies, New York, NY, USA, 2013, pp. 163–174.

[50] H. H. Cho, C. F. Lai, T. K. Shih, and H. C. Chao, "Integration of SDR and SDN for 5G," IEEE Access, vol. 2, pp. 1196–1204, 2014.

[51] I. F. Akyildiz, P. Wang, and S.-C. Lin, "SoftAir: A software-defined networking architecture for 5G wireless systems," Comput. Networks, vol. 85, pp. 1–18, 2015.

[52] S. Namal, I. Ahmad, S. Saud, M. Jokinen, and A. Gurtov, "Implementation of OpenFlow-based cognitive radio network architecture: SDN&R," Wirel. Networks, vol. 22, no. 2, pp. 663–677, 2016.

[53] M. Boussard, D. T. Bui, L. Ciavaglia, R. Douville, M. Le Pallec, N. Le Sauze, and F. Santoro, "Software-Defined LANs for Interconnected Smart Environment," in Teletraffic Congress (ITC 27), 2015 27th International, 2015, pp. 219–227.

[54] A. Darabseh, M. Al-Ayyoub, Y. Jararweh, E. Benkhelifa, M. Vouk, and A. Rindos, "SDStorage: A Software Defined Storage Experimental Framework," in Cloud Engineering (IC2E), 2015 IEEE International Conference on, 2015, pp. 341–346.

[55] X. Xiong, L. Hou, K. Zheng, W. Xiang, M. S. Hossain, and S. M. M. Rahman, "SMDP-Based Radio Resource Allocation Scheme in Software-Defined Internet of Things Networks," IEEE Sensors J., vol. PP, no. 99, pp. 1–1, 2016.

[56] A. A. Corici, R. Shrestha, G. Carella, A. Elmangoush, R. Steinke, and T. Magedanz, "A solution for provisioning reliable M2M infrastructures using SDN and device management," in Information and Communication Technology (ICoICT), 2015 3rd International Conference on, 2015, pp. 81–86.

[57] A. Dunkels, B. Gronvall, and T. Voigt, "Contiki - A Lightweight and Flexible Operating System for Tiny Networked Sensors," in Proceedings of the 29th Annual IEEE International Conference on Local Computer Networks, Washington, DC, USA, 2004, pp. 455–462.

[58] E. Baccelli, O. Hahm, M. Gunes, M. Wahlisch, and T. C. Schmidt, "RIOT OS: Towards an OS for the Internet of Things," in Computer Communications Workshops (INFOCOM WKSHPS), 2013 IEEE Conference on, 2013, pp. 79–80.

[59] P. Levis, Madden, S., Polastre, J., Szewczyk, R., Whitehouse, K., Woo, A., and Culler, D., "TinyOS: An Operating System for Sensor Networks," in Ambient Intelligence, W. Weber, J. M. Rabaey, and E. Aarts, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2005, pp. 115–148.

[60] Q. Cao, T. Abdelzaher, J. Stankovic, and T. He, "The LiteOS Operating System: Towards Unix-Like Abstractions for Wireless Sensor Networks," in Proceedings of the 7th International Conference on Information Processing in Sensor Networks, Washington, DC, USA, 2008, pp. 233–244.

[61] T. Koponen, Casado, M., Gude, N., Stribling, J., Poutievski, L., Zhu, M.and Shenker, S. "Onix: A Distributed Control Platform for Large-scale Production Networks," in In Proc. OSDI, 2010.

[62] E. Ng, "Maestro: A system for scalable OpenFlow control," Rice Univ., 2010.

[63] S. Costanzo, L. Galluccio, G. Morabito, and S. Palazzo, "Software Defined Wireless Networks: Unbridling SDNs," in 2012 European Workshop on Software Defined Networking, 2012, pp. 1–6.

[64] R. Sherwood, G. Gibb, K. K. Yap, G. Appenzeller, M. Casado, N. McKeown, and Parulkar, G "Flowvisor: A network virtualization layer," OpenFlow Switch Consort. Tech Rep, pp. 1–13, 2009.

[65] L. Galluccio, S. Milardo, G. Morabito, and S. Palazzo, "SDN-WISE: Design, prototyping and experimentation of a stateful SDN solution for WIreless Sensor networks," in 2015 IEEE Conference on Computer Communications (INFOCOM), 2015, pp. 513–521.

[66] A. C. G. Anadiotis, L. Galluccio, S. Milardo, G. Morabito, and S. Palazzo, "Towards a software-defined Network Operating System for the IoT," in Internet of Things (WF-IoT), 2015 IEEE 2nd World Forum on, 2015, pp. 579–584.

[67] P. Berde, M. Gerola, J. Hart, Y. Higuchi, M. Kobayashi, T. Koide, and G. Parulkar, "ONOS: towards an open, distributed SDN OS," in Proceedings of the third workshop on Hot topics in software defined networking, 2014, pp. 1–6.

[68] K. S. Sahoo, B. Sahoo, and A. Panda, "A secure SDN framework for IoT," in 2015 International Conference on Man and Machine Interfacing (MAMI), 2015, pp. 1–4.

[69] C. Gonzalez, S. M. Charfadine, O. Flauzac, and F. Nolot, "SDN-based security framework for the IoT in distributed grid," in 2016 International Multidisciplinary Conference on Computer and Energy Science (SpliTech), 2016, pp. 1–5.

[70] "OpFlex: An Open Policy Protocol White Paper," Cisco.

[71] S. Chakrabarty and D. W. Engels, "A secure IoT architecture for Smart Cities," in 2016 13th IEEE Annual Consumer Communications & Networking Conference (CCNC), 2016, pp. 812–813.

[72] O. Flauzac, C. González, A. Hachani, and F. Nolot, "SDN Based Architecture for IoT and Improvement of the Security," in Advanced Information Networking and Applications Workshops (WAINA), 2015 IEEE 29th International Conference on, 2015, pp. 688–693.

[73] F. Olivier, G. Carlos, and N. Florent, "New Security Architecture for IoT Network," Procedia Comput. Sci., vol. 52, pp. 1028–1033, 2015.

[74] R. E. Balfour, "Building the Internet of Everything (IoE) for first responders," in Systems, Applications and Technology Conference (LISAT), 2015 IEEE Long Island, 2015, pp. 1–6.

[75] C. Vandana, "Security improvement in IoT based on Software Defined Networking (SDN)."

[76] A. Ukil, J. Sen, and S. Koilakonda, "Embedded security for Internet of Things," in Emerging Trends and Applications in Computer Science (NCETACS), 2011 2nd National Conference on, 2011, pp. 1–6.

[77] R. Skowyra, S. Bahargam, and A. Bestavros, "Software-defined ids for securing embedded mobile devices," in High-Performance Extreme Computing Conference (HPEC), 2013 IEEE, 2013, pp. 1–7

Modified Random Forest Approach for Resource Allocation in 5G Network

Parnika De
Department of Computer Engineering and Applications,
National Institute of Technical Teachers Training and Research, Bhopal, India

Shailendra Singh, Senior Member IEEE
Department of Computer Engineering and Applications,
National Institute of Technical Teachers Training and Research, Bhopal, India

***Abstract*—According to annual visual network index (VNI) report by the year 2020, 4G will reach its maturity and incremental approach will not meet demand. Only way is to switch to newer generation of mobile technology called as 5G. Resource allocation is critical problem that impact 5G Network operation critically. Timely and accurate assessment of underutilized bandwidth to primary user is necessary in order to utilize it efficiently for increasing network efficiency. This paper presents a decision making system at Fusion center using modified Random Forest. Modified Random Forest is first trained using Database accumulated by measuring different network parameters and can take decision on allocation of resources. The Random Forest is retrained after fixed time interval, considering dynamic nature of network. We also test its performance in comparison with existing AND/OR logic decision logic at Fusion Center**

Keywords—5G; Cognitive Radio; Clustering; Fusion Centre; Random Forest

I. INTRODUCTION

History of modern communication starts with inception of electrical telegraph system, which uses Morse code for communication between two distant locations.

Due to limitation in existing technology i.e. only data can be sent in form of Morse code and at other end person should be there to decode Morse code information as well as sending long messages is not recommended. In an independent attempt made by Graham Bell in 1837, to transmit voice signal from one location to another, he invented a device which he named 'Telephone'. This communication technology revolutionize whole scenario and remain most popular means for coming century. In year 1948, Professor Shannon presented a paper on "*A Mathematical Theory of Communication*" [1]. Idea presented in the paper was way ahead of its time, he suggested use of 0's and 1's for communication. Evolving itself and surrounding is nature of Human being, in the early 90's internet came into existence as another mode for sharing and communicating information. In earlier times it was limited to wired communication. It limits users to use internet at static location and hence lack feature of mobility. Being a Human, we want our machines to act like us, mobility inspired researcher and hence result is mobile revolution. Like any other technology it is also continuously evolving itself i.e. 1G, 2G, 3G, 4G (currently deployed) and 5G (under Research domain) [2][3], [4].

Question comes why we need another generation of mobile Technology?

It is stated by annual visual network index (VNI) report released by CISCO [5] that data explosion in wireless communication will continue in coming years also. VNI report also stated 4G Network will unable to handle network load with incremental approach, it will also reach to its maturity by the year 2020.

For introduction of 5G Network Technology in reality following requirements must be fulfilled [5]–[7]:

- *Data Rate*

Data rate is amount of data transfer per second per unit area. Considering 5G into scenario it would be 1000 times more than current 4G Network.

- *Latency*

4G has latency of 15ms, due to future demand of services like online gaming and virtual reality, latency in 5G Network should not exceed 1ms.

- *Energy and cost*

It is stated by researcher in 5G Network cost and energy consumption will reduce. Energy and cost is measured in Network with Joules/bit and cost/bit will fall up to 100 fold in 5G Network.

- *Battery*

Conserving battery life in Network is main concern for 5G Network consumption of battery life is reduce up to 10 times than the existing 4G Network.

If above requirement gets fulfilled, 5G will become reality which can offer fast connectivity without any constraint on all available devices.

Following techniques enable 5G dream into reality:

- *Densification of Network*

Densification of network is done through deployment of large number of small cell; decreasing cell size is big challenge for researchers. Most recent development is in japan where cell spacing is reduced to $1/10^{th}$ of square kilometer.

- *Massive MIMO*

MIMO was introduced in year 2006. It consider spatial dimension of communication, if multiple antennas are available at the Base Station. These techniques harness the multiplexing feature to get better results. Multi user-MIMO is included into 3GPP LTE- advanced standard, still higher capacity is yet to achieve. In VLM-MIMO, number of antennas per cell is larger than number of user, result in many desirable features.

- *Millimeter (mm-wave) signals*

In search of free frequency band researcher found 30-300 GHz are free available bandwidths. Total available spectrum in this range is 200 times greater than current used frequency i.e. 3GHz.

- *Direct Device to Device (D2D)*

It is exchange of data between mobile devices without Base Station in between, in result it reduce load over network devices which has to handle hundreds and thousands of requests simultaneously. It has to deal with heterogeneity of network, to support heterogeneity multiple protocols has been decided.

- *Full duplex wireless*

Full duplex enable both side to transmit data simultaneously in same frequency band, it has numerous benefit it increase physical layer capacity up to twice and also improves latency and security at physical layer.

A. *Network Architecture*

In Heterogeneous Network environment where network is flooded with numerous devices, architectures and protocols standardization of network is very important. Currently 3GPP is finalizing LTE-Rel-12 (third release of LTE-advanced family) [5], [7], [8]. It is expected that standardization of 5G will not be finalized until Rel-14 and Rel-15. Currently there are numerous architecture and proposed protocol existing for the 5G network. Many organizations are currently working to make advancement in 5G Network Technology at their own level with collaboration of different universities and Telecom industrial funding. Such as METIS project in Europe, IMT-2020 of China, 5G forum of Korea and ADWICS of Japan. One of the most acclaimed architecture for futuristic 5G Technology is to divide Network on the basis of rights over Bandwidth. Rights over Bandwidth is allocated to two class of users named as Primary Users and Secondary Users [9]–[14]. Primary users are class of users who has basic rights over bandwidth and Secondary user get control on Bandwidth whenever Primary Users are ideal.

The platform is set for the working of Cognitive Radio, in which Secondary User has to intelligently detect which bandwidth is under use and which is not. This optimizes use of available bandwidth while minimizing interference to other Secondary users and Primary Users.

Main challenge is how to detect the vacant Bandwidth?

Solution comes from hybridizing Basic Network architecture for 5G with one of the proposed architecture for wireless sensor Network.

Densification of Network is done through massive deployment of cell in the heterogeneous Network Environment, with the advancement in cell Technology cell size is getting reduced and currently it is $1/10^{th}$ square kilometer in Japan [5]. Sometimes these small cell is known from the name Base Station [10], [13]–[15], work of these Base Station is to collect signals from different Mobile Devices active in its coverage area. There are many optimization techniques suggested by Researchers for collecting data from different mobile devices.

Embedding the concept of Wireless Sensor Network in 5g network environment, base station enables communication between each other from a multi-hop network. Cost of transmitting is higher than computation [5], [16], [17], because of this base Station are organized into Clusters.

In the Cluster environment of the Base station, data is collected by central processing center which is a specialized device for entertaining request made by base station for resource allocation.

In the cluster environment, each cluster select its cluster head through many proposed methods by researchers [18], [19].

There are two basic scheme of clustering Base Stations:

- Single Layer Clustering

In this clustering approach [16], [17], all Base station are clustered having each cluster its own Cluster Head. Each cluster Head directly transmits its signal to data processing center also known as fusion center in 5G Network.

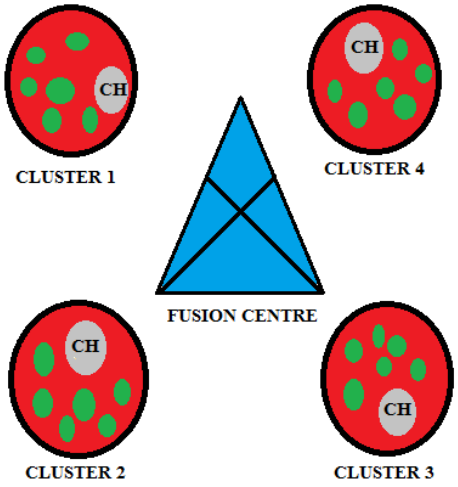


Fig. 1. Single layer clustering

- Hierarchical clustering

As suggested in [16], [17], authors assumed there j levels of clustering. Level 1 is at lowest level and level j is highest level. In hierarchical environment each lower id cluster head transmits data to its immediate upper layer cluster head and so on.

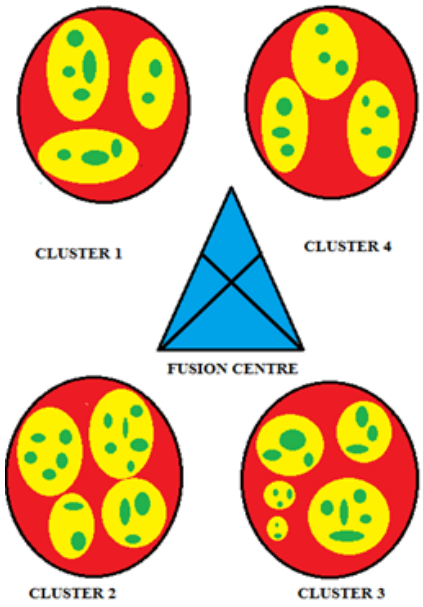


Fig. 2. Hierarchical clustering

II. Related Work

A. 5G and Cognitive Radio

Before idea of wireless communication network, wired communication was very much popular up to late 90's. dawn of mobile communication technology ignited light for rapid evolution in mobile generation like 1G,2G,3G and 4G [2], [3], [26], [27]. Researchers contributed a lot for development in different sphere of network. These attribute are performance, architecture and cost. Being 5G as futuristic technology it has many challenges and requirements issues like spectrum allocation, speed, cost and data traffic [7], [8], [28]. Challenges can be conquered if network has strong architecture in this regard a lot of research work has been done. Most of researchers have suggested implementation of massive MIMO, application of millimeter wave, device to device communication and embedding cognitive radio with 5G, full duplex communication, ultra dense network using small cell and inference management [5], [6]. Continuous works are carrying on by researchers, for each of the suggested improvement parameter mentioned above[20], [21], [22].

Since spectrum being most vital resource in network but spectrum allocation and sensing is one of the challenging task for research community where reliability and accuracy matters a lot. Researchers suggested many models co-operative spectrum sensing [29]–[31] is very much popular in which two type of users, primary user and secondary user. Both collaborate with each other for spectrum sharing with the help of active and inactive phase. For sensing spectrum, an intelligent software radio was suggested also known as cognitive radio [9]–[13], [32]–[51]. In 2003, Freidreich K. Jondral et al. paper [52] on smart radio later known from name cognitive radio. Many of research work are available on cognitive radio explaining, modifying and implementing cognitive radio technology [9]–[13], [32]–[51].

Cognitive radio is an "intelligent communication system that is aware of its surrounding environment, and uses the methodology of understanding by building to learn from environment and adapt its internal states to statistical variation in the incoming radio frequency stimuli by making certain changes in operating parameter in real time, with two primary objectives in mind: highly reliable communication and whenever needed efficient utilization of radio spectrum." Haykins. For implementation of cognitive radio a tool is needed, machine learning [31] is pioneer candidate among all of them because it learn from past data and derive knowledge base from it and able to take decision with any manual help. In 2013, paper [31] named "*Machine Learning Techniques for Cooperative Spectrum Sensing in Cognitive Radio Networks*" which discuss all major machine learning algorithms and how they can be used for spectrum sensing in cognitive radio network. After publishing of this paper research community showed interest for using machine learning algorithm for spectrum sensing. Machine learning can be implemented in many other area of network and much more can be achieved through it in network scenario.

B. Random Forest

In year 2001, Leo Breiman, a statistician identifies the problems in existing machine learning techniques [53]. In earlier tree approach of machine learning data set is not evenly distributed lead to imbalance of data. Imbalanced data set performance is poor with the classification, this lead to miss classification and error in the training phase. Leo Breiman in suggested a new machine learning algorithm to improve the classification of diverse data, it used his own 'bagging' idea [23] and Ho and Amit and Geman's random selection [54] to construct number of Decision Tree with control variance. He suggested data set were collected and then divided into two or more subset of data, where one or more data set used as learner and remaining is used for test purpose.

Many researchers got attracted towards Random Forest approach of handling data set and started working on different attributes of Random Forest like features, concepts, analysis and modification of the proposed model of Random Forest algorithm. Research works going on in the field of Random Forest can be broadly classified into three categories:

- Research in Random Forest Improving accuracy
- Improving performance (performance deals with reducing time for learning and classification)
- Exploration of new domain for application of Random Forest

Number of tree generated in the Random Forest is the challenge for the Researchers because it consumes extra space in memory and also increases run time of the algorithm, solution comes in two forms parallel computing and Pruning of Random Forest. Research work [25] done under SPRINT which uses 128 process and speed up performance up to 50 times over the serial code. Basic problem with this approach is that it takes only time complexity under consideration leaving the space complexity. Because of this pruning approach is required. Basically two approaches are followed one is static and other one is dynamic approach [25]. Static approach

follow overproduce and choose characteristic [25], [55] because decision trees are first overproduce in Random Forest to pre decided number and then in the choose strategy best decision tree got selected. Whereas dynamic pruning do not have over produce phase [25] . Saving more time when comparison to static approach but unfortunately it is hard to implement because of this researchers are also not showing interest for this approach. Research work done under static pruning approach fall in to three majority categories:

- Weighted voting method
- Ranking Based method
- Search Based method

In one pioneer work [25]of static pruning, genetic algorithm is used to select most optimal candidate from pool of Decision Tree. Other work uses elimination [25], [55] of similar Decision Tree if their output class and accuracy are same then keep single copy of Tree eliminating others. Dynamic pruning require help of statistics and probability along with nature inspired algorithm to get better results. In one approach [25] authors has tried to model dynamic pruning approach with the help of eight degree mathematical equation.

C. 5G in Machine Learning

Machine learning algorithms are either classifiers or regression. They take the help of past data and accumulate knowledge from it and also take decision. It has been demand from long time to implement machine learning algorithm for network scenario and utilize strength of it for increasing reliability and accuracy of 5G. In network, there are various place where machine learning algorithm can be used. For clustering of mobile node, various clustering algorithm can be used. Researchers also pointed out various suitable algorithms for the clustering in network scenario. Hierarchical and weighted clustering algorithm is generally suggested by researcher [18], [19], [56], [57]. Some researchers also suggest modified version of both the algorithm to achieve better output. Other place where machine learning algorithm can be used is for spectrum sensing. In paper [31] author has suggested many such algorithm for sensing vacant spectrum sensing. Further extension of spectrum sensing is spectrum allocation, machine learning algorithms are very much capable of handling both load without manual help i.e. totally autonomous system. In some of the paper researchers also fuse probability and machine learning for spectrum sensing. Nature inspired algorithm can also be used for implementation in 5G. One author has already used genetic algorithm [58], [59], he matched genetic parameter with network parameter, and with the help of it he has predicted variation in network parameters.

III. Assumptions and Gaps Identified

A. Assumptions

5G network is very dynamic in nature, data collected from mobile nodes and implementation of decision making mechanism at fusion center has to solve certain issues.

- The proposed method assumes there should be no link failure between CH and FC.
- Fair transmission of data between FC and CH without any delay.
- No assumption has been taken about working environment of FC.

B. Research Gap

The 5G technology is futuristic, for implementation of 5G into reality, there are numerous issues has to be conquer. One of the most challenging issue which is still untouched by researcher is taking most optimal decision at Fusion Center (FC). In Network spectrum allocation is a crucial issue, reason is limited bandwidth whereas number of users is always more and increasing exponentially. FC is totally responsible for bandwidth allocation, a single mistake in spectrum allocation will cost a lot to the Network. Following consequences may arise due wrong allocation of spectrum by the Fusion Center:

1) *Wastage of bandwidth*
2) *It may lead to monopolization of spectrum.*
3) *It may cause deadlock to the Network.*

IV. Methodology

A. Proposed Method

Mobile nodes in the 5G Network are of two type primary user and secondary user. All mobile node send its information to the Cluster Head. Information send to the CH consist of attribute and its value for particular mobile node. Attributes values explain about status of mobile node in the Network, these attribute are battery life of mobile node, distance of mobile node from the CH, weather it is primary user or secondary user, signal to noise ratio, if node is primary user weather it is using Bandwidth or not and many other attributes [16]. Each cluster head send these information of mobile nodes to the fusion center, a central hub equipped with all essential hardware and Software components. Architecture of Fusion center must consist of following component:

- A receiver antenna for collecting all information from different Cluster Head.

- Processing unit either single unit or multiple unit for parallel processing
- Database storage
- Transmitter unit for broadcasting output of processing unit to the cluster Heads.
- Display unit for monitoring the output of processing unit
- Input unit for feeding algorithms and different variables to the processing system.

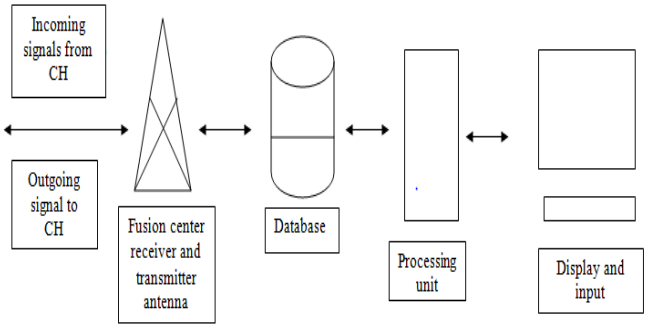


Fig. 3. Working of Random Forest Algorithm

B. Working of Random Forest Algorithm

All manipulation work on the Database is done here at the Processing Unit. Processing unit fetch data set from Database, there are two type of Database, static and dynamic. Static database is source for training Random Forest it contain tables in form of rows and columns where rows represent past recorded information about each mobile node and column represent attribute values along with decision attribute value for the table. This decision value illustrate about under certain values of attribute which cluster has got the spectrum allocation, this help Random Forest algorithm in learning process. Dynamic Database is continuously coming information from cluster head which is treated as request for spectrum allocation, after taking Decision making for request particular table is stored into static database for further learning of Random forest which is repeated after fixed time interval considering dynamic nature of network where different cluster and mobile nodes are getting detached and attached to network continuously, because of this reason we need to re train our Random Forest after fixed time interval, so that it can take most optimal and accurate decisions. In future work of paper we will also consider the problem of re training the Random Forest in real time duration with the dynamic database. Current work is based on decision making at fusion center using static database which reduces time complexity of random forest using static pruning methodology applying concept of clustering (K-mean) [60] of Decision Tree, which is illustrated through following steps:

1. *Retrieve Data set from database in the form of table.*
2. *Apply bagging or bootstrapping approach on the data set, it divide the data set into different subset of data set with replacement of rows.*
3. *Build Decision Tree on each of the sub samples, applying random sub sampling of attributes.*
4. *Measure accuracy of each of the Decision Tree.*
5. *Name each of the Decision Tree*
6. *Store each of the Decision Tree accuracy along with its respective names.*
7. *Apply K-mean clustering algorithm for clustering Decision Tree on the basis of accuracy measure.*
8. *Calculate centroid of each cluster*
9. *Apply following formula:*

Let 'S' be the sum of all cluster's centroid.

p_i be the accuracy of ith cluster's centroid.

'M' be the total number of clusters.

w_i be the weight of ith cluster

N be the total number of Decision Tree in the Random Forest.

s_i is the number of Decision tree selected from ith cluster.

a. Below formula illustrate about sum of all cluster's centroid.

$$S = \sum_{i=1}^{i=M} P_i \quad (1)$$

b. Repeat this step M number of times

$$w_i = \frac{p_i}{s} \quad (2)$$

c. Repeat this step M number of times

$$s_i = w_i * N \quad (3)$$

If s_i value is decimal take the floor value, if total number of tree in the particular cluster is less than s_i value select all Decision Tree from cluster.

10. Use voting method for class selection i.e. selection of cluster for spectrum allocation (this is mobile node cluster do not confuse with DT cluster)
11. Allocate spectrum to most voted cluster.

As par we have gone through all literature survey, this approach is not implemented by any researcher for designing decision making system at fusion center. There exist no research work done until yet for smart decision making for fusion center, currently AND/OR logic is available to take decision at fusion center.in the next section of paper we will show experimental results comparing our approach with traditional AND/OR logic approach. After going through a lot of research paper Random Forest seems to be most suitable in deployment of 5G Network [61]–[64], reason is accuracy and capability to become classifier of future generation. Since, 5G is also futuristic technology it require a strong and more accurate classifier. Main concern of Researchers of Random Forest is to enhance performance if we want to use it under network condition. Paper contributed to decrease execution time by using clustering approach and reduce the time for classification. We will also do a comparison of our proposed algorithm with traditional Breiman Random Forest model [53] in very next section and check which algorithm requires less time. To have clear understanding of concepts in proposed work a diagram is given below which explain each steps clearly.

V. Result Analysis

5G is not implemented yet, due to this reason for the result analysis we have to collected data from different resources. Many research organization is currently working on this futuristic technology we want to thank Nokia Telecommunication, Samsung research laboratories and

METIS project for their support in my experimental work [16], [17], [28]. It is due to them we are able to gather all key features and their values. For assembling these data provide we have taken the help of different researcher working in this field, there research work [14], [30], [65]–[68], [59]has guided us to assemble these value.

For our experiment purpose, we have created datasets representing small database of fusion center where each row represent information of single mobile node and column represent attribute values. Experiment is conducted on dataset having 153 rows and 7 columns for static database. Since, column represent attributes these attributes are battery life (b_life) which explains about remaining battery life of mobile node, distance from cluster head (dist_CH) as the name describe it describe about mobile node distance from its cluster head, type of mobile node (mob_node) since we know that there two type of user on the bases of spectrum rights it explains about weither it is primary or secondary user, vacant bandwidth (vac_band) this is important feature for the primary user because on the bases of this field, it is understood primary user or secondary user is utilizing bandwidth allocated to them, signal strength (sig_strenth) signal strength is measure on the scale of dBm , cluster identity (c_id) describes about mobile node belong to which cluster many mobile node can have same cluster id and output (o_put) tells about under certain condition spectrum is allocated to which cluster. Static database is used for training multiple Decision trees. Portion of dataset is shown below:

TABLE I. DATASET COLLECTED FROM MOBILE NODE AND CLUSTER

Identity of mobile node	B_ life	dist_ CH	mob_ node	vac_ band	sig_ strength	c_ id	o_ put
10	51	212	1	1	-44.3	4	4
21	22	110	0	0	-30.2	12	16
18	18	143	0	1	-31.6	9	9
144	94	361	1	1	-50.9	7	7
32	12	215	1	0	-42.7	11	4
56	62	311	0	1	-46.2	5	5
91	97	165	1	0	-32.1	6	8

From the above table it is easy to deduce that if primary user (represented by 1) and secondary user (represented by 0) is using its allocated bandwidth, bandwidth cannot be allocated to another user or cluster and represented in vac_band column with 1 showing user is using allocated bandwidth. In this case both c_id and o_put both column values are same for particular rows. These values in the table can be dropped because cannot be used for m training classifiers (Random Forest). This result in reduce in the table size.

TABLE II. FINAL DATA SET USED FOR IMPLEMENTATION

Identity of mobile node	b_ life	dist_ CH	mob_ node	vac_ band	sig_ strength	c_ id	o_ put
21	22	110	0	0	-30.2	12	16
32	12	215	1	0	-42.7	11	4
91	97	165	1	0	-32.1	6	8

Result analysis cannot be possible without implementation of proposed algorithm and we also need a tool for comparison of existing approach with proposed approach. For this purpose, R seems to be the best choice because of its plug and play libraries and ease of use. For execution of algorithm all parameters were kept same in hardware and software manner. Both the algorithm were run on xenon processor with clock speed 3.45Ghz.All results shown below in two sub-section is implemented in latest version of R i.e. R 3.2.3.

A. *Modified Random Forest Algoritm*

Our purpose for proposed algorithm is to enhance performance (reduce execution time of Random Forest algorithm) without losing accuracy measure. Experiment is conducted on three data set i.e. modified Random Forest and Breiman Random Forest is evaluated using three data set on back ground of accuracy and Performance. All three data sets used in the experiment are shown in appendix section. We have also done portioning of each of data set into three parts those parts are training Data set, validation data set and testing data set. Data set with same portioning value is applied to both algorithms.

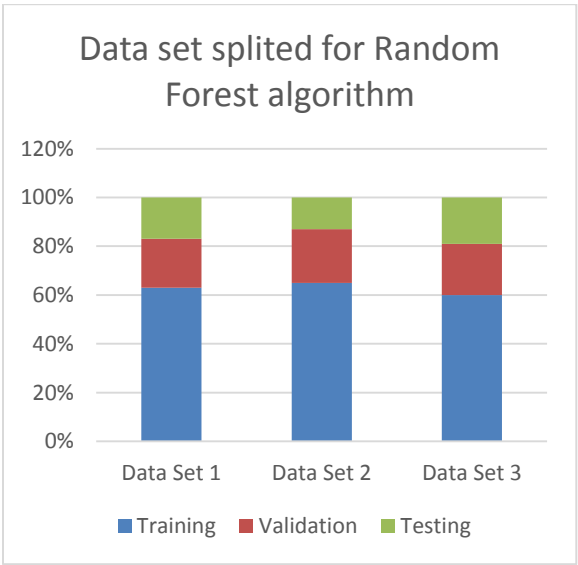


Fig. 4. Bar Graph of data set splitted for Random Forest

Through bar graph it is illustrated in data set 1, 63% of data set rows are allocated for training purpose of Random Forest, 20% for the validation and remaining 17% is allocated for testing. Similarly in the data set 2, 65%, 22% and 13% is allocated for training, validation and testing purpose respectively. Similar splitting of data set is done in the data set 3 also with 60%, 21% and 19% for training, validation and testing purpose.

After running data set1 data set 2 and data set 3 in Breiman algorithm and Modified RF, for testing results two parameters are accuracy and performance.

Evaluating on the basis of accuracy, it is clear from results shown with the help of graphs accuracy of proposed modified Random Forest algorithm is very much near to Breiman's Random Forest Algorithm.

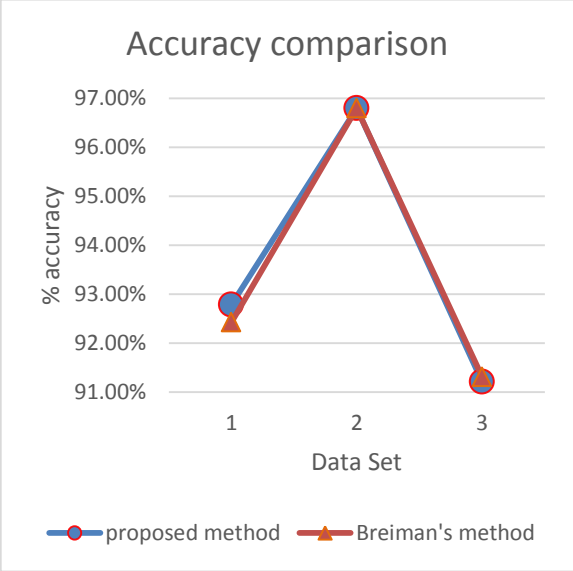


Fig. 5. Accuracy Comparison of actual method and the suggested method

Another parameter for comparison is the time complexity, both the algorithm requires time in mille seconds using R editor for implementation. From the results produced after execution of both algorithms it is clear that proposed modified Random Forest approach needs less time for every data set taken for experiment. Hence considering performance factor proposed algorithm is more impact full than existing Breiman's Random Forest. Analysis of results can be done with the help of two graphs given below:

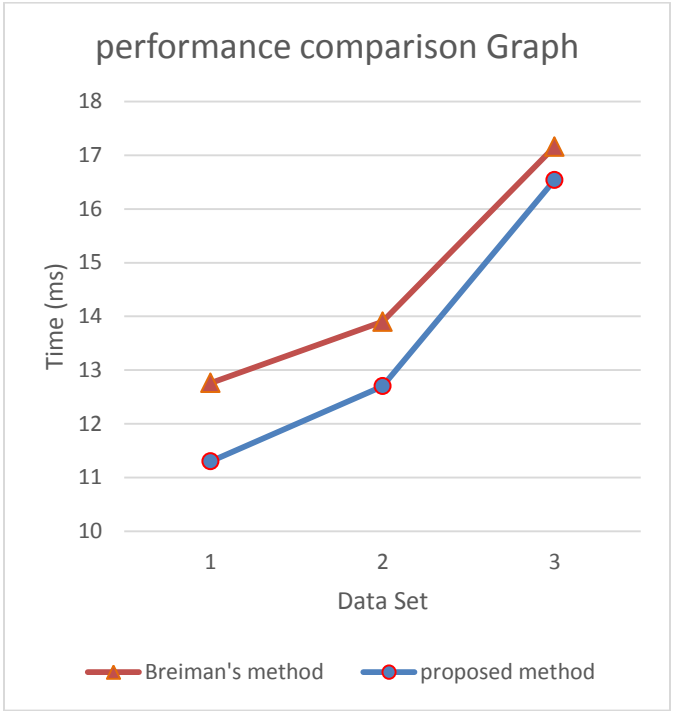


Fig. 6. Performance comparison of actual Breimann's Random Forest and the modified Random Forest

B. *Modified Random Forest Algorithm at the Fusion Centre*

New algorithm is used to increase accuracy of spectrum allocation. So that wise utilization of spectrum can take place. For this purpose, machine learning technique is used because it take decision based on prior knowledge accumulation. Random Forest is most optimal among all machine learning techniques considering accuracy as a parameter under network load condition. Up to know statement was just a hypothesis, results are very much promising. Comparing results of the modified Random Forest algorithm with existing AND/OR logic approach of decision making at fusion center, we can see significant improvement in accuracy for spectrum allocation and spectrum allocation time to the cluster head also decreases and perform faster than AND/OR logic. Experimental result can be analyzed from graph shown below, all three data set is used for experimental output generation.

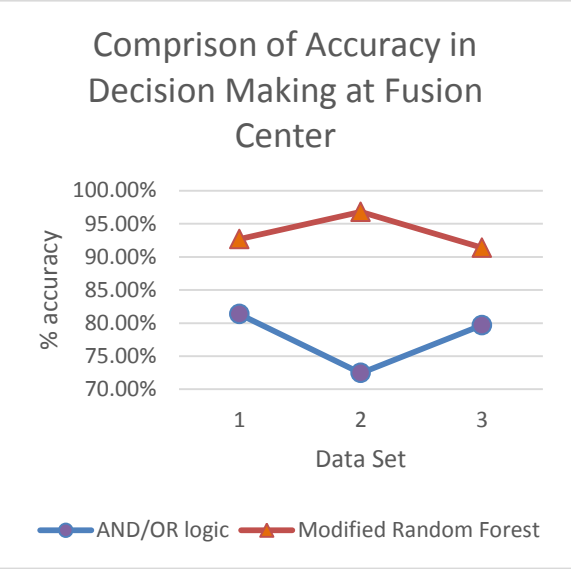


Fig. 7. Comparison of accuracy in decision making at the fusion centre

VI. Conclusion

This paper gives a new algorithm, Modified Random Forest Algorithm for 5G. Due to the use of this algorithm we can see that there is an improvement in the performance of the 5G network for resource allocation.

In future the work can be extended to the full 5G network rather than only for resource allocation.

References

[1] "A Mathematical Theory of Communication," vol. 27, no. July 1928, pp. 379–423, 1948.

[2] P. De, "A survey on 5G with Cognitive Radio Technology."

[3] A. Gohil, H. Modi, and S. K. Patel, "5G Technology of Mobile Communication : A Survey," pp. 288–292, 2013.

[4] C. Paper, "5G technology of mobile communication : A survey 5G Technology of Mobile Communication : A Survey," no. October, 2015.

[5] J. G. J. G. Andrews, S. Buzzi, W. Choi, S. V. S. V. Hanly, A. Lozano, A. C. K. A. C. K. Soong, and J. C. J. C. Zhang, "What will 5G be?," *IEEE J. Sel. Areas Commun.*, vol. 32, no. 6, pp. 1065–1082, 2014.

[6] S. Talwar, D. Choudhury, K. Dimou, E. Aryafar, B. Bangerter, and K. Stewart, "Enabling Technologies and Architectures for 5G Wireless," 2014.

[7] S. Chen and J. Zhao, "The requirements, challenges, and technologies for 5G of terrestrial mobile telecommunication," *IEEE Commun. Mag.*, vol. 52, no. 5, pp. 36–43, 2014.

[8] J. M. Iii, J. Guerci, J. Reed, Y. Yao, Y. Chen, and T. C. Clancy, "Accelerating 5G QoE via Public-Private Spectrum Sharing," no. May, pp. 77–85, 2014.

[9] "Cognitive Radio : Making Software Radios More Personal."

[10] S. Stotas and A. Nallanathan, "On the Throughput and Spectrum Sensing Enhancement of Opportunistic Spectrum Access Cognitive Radio Networks," vol. 11, no. 1, pp. 97–107, 2012.

[11] S. Wang and H. Zheng, "A RESOURCE MANAGEMENT DESIGN FOR COGNITIVE RADIO AD HOC NETWORKS," pp. 1–7, 2009.

[12] A. Jovii and P. Viswanath, "Cognitive radio: An information-theoretic perspective," *IEEE Trans. Inf. Theory*, vol. 55, no. 9, pp. 3945–3958, 2009.

[13] S. Stotas and A. Nallanathan, "Enhancing the Capacity of Spectrum Sharing Cognitive Radio Networks," *October*, vol. 60, no. 8, pp. 3768–3779, 2011.

[14] C. R. W. Thomas, "Cognitive Networks," 2007.

[15] S. Teli, "International Journal of Advanced Research in Computer Science and Software Engineering A Survey on Decision Tree Based Approaches in Data Mining," vol. 5, no. 4, pp. 613–617, 2015.

[16] S. Fletcher and N. E. C. Telecom, "Cellular Architecture and Key Technologies for 5G Wireless Communication Networks," no. February, pp. 122–130, 2014.

[17] N. Networks, "Network architecture for the 5G era."

[18] M. R. Brust, A. Andronache, and S. Rothkugel, "WACA : A Hierarchical Weighted Clustering Algorithm optimized for Mobile Hybrid Networks."

[19] Y. Wang and F. S. Bao, "An Entropy-based Weighted Clustering Algorithm and Its Optimization for Ad hoc Networks," no. WiMOB, 2007.

[20] M. Learning, K. A. Publishers, A. C. Sciences, and R. August, "Induction of Decision Trees," pp. 81–106, 2007.

[21] W. Liu, S. Chawla, D. A. Cieslak, and N. V Chawla, "A Robust Decision Tree Algorithm for Imbalanced Data Sets C4 . 5 on Balanced Data When data is balanced , C4 . 5 gives a good boundary between the two," 2010.

[22] A. M. Prasad, L. R. Iverson, and A. Liaw, "Newer Classification and Regression Tree Techniques : Bagging and Random Forests for Ecological Prediction," no. December 2004, pp. 181–199, 2006.

[23] L. Breiman, "Bagging Predictors," no. 421, 1994.

[24] M. Fern and E. Cernadas, "Do we Need Hundreds of Classifiers to Solve Real World Classification Problems ?," vol. 15, pp. 3133–3181, 2014.

[25] V. Y. Kulkarni and P. K. Sinha, "Pruning of random forest classifiers: A survey and future directions," *Proc. - 2012 Int. Conf. Data Sci. Eng. ICDSE 2012*, pp. 64–68, 2012.

[26] M. R. Bhalla and A. V. Bhalla, "Generations of mobile wireless technology: a survey," *Int. J. Comput. Appl.*, vol. 5, no. 4, p. 7, 2010.

[27] I. C. Surveys, "A Survey of Spectrum Sensing Algorithms for Cognitive Radio Applications," vol. 11, no. 1, pp. 116–130, 2009.

[28] A. Osseiran, F. Boccardi, V. Braun, K. Kusume, P. Marsch, M. Maternia, O. Queseth, M. Schellmann, H. Schotten, H. Taoka, H. Tullberg, and M. A. Uusitalo, "Scenarios for 5G Mobile and Wireless Communications : The Vision of the METIS Project," no. May, pp. 26–35, 2014.

[29] F. Ge, Q. Chen, Y. Wang, C. W. Bostian, T. W. Rondeau, and B. Le, "Cognitive Radio : From Spectrum Sharing to Adaptive Learning and Reconfiguration," 2008.

[30] C. R. Networks, "Transactions Letters," vol. 7, no. 12, pp. 4761–4766, 2008.

[31] K. M. Thilina, N. Saquib, and E. Hossain, "Machine Learning Techniques for Cooperative Spectrum Sensing in Cognitive Radio Networks," *IEEE J. Sel. Areas Commun.*, vol. 31, no. 11, pp. 2209–2221, 2013.

[32] Z. Shu, G. Wang, X. Tian, D. Shen, K. Pham, E. Blasch, and G. Chen, "RESOURCE ALLOCATION IN UNDERLAY COGNITIVE RADIO SATCOM," pp. 1–8, 2015.

[33] E. Rajo-iglesias, "Cognitive-Radio and Antenna Functionalities: A Tutorial," vol. 56, no. 1, pp. 231–243.

[34] S. Pandit, S. Member, and G. Singh, "Backoff Algorithm in Cognitive Radio MAC Protocol for Throughput Enhancement," vol. 64, no. 5, pp. 1991–2000, 2015.

[35] L. Gavrilovska, S. Member, D. Denkovski, S. Member, V. Rakovic, S. Member, and M. Angjelichinoski, "Medium Access Control Protocols in Cognitive Radio Networks : Overview and General Classification," vol. 16, no. 4, pp. 2092–2124, 2014.

[36] R. K. Mclean, S. Member, M. D. Silvius, K. M. Hopkinson, S. Member, B. N. Flatley, S. Member, E. S. Hennessey, S. Member, C. C. Medve, S. Member, J. J. Thompson, S. Member, M. R. Tolson, C. V Dalton, and S. Member, "An Architecture for Coexistence with Multiple Users in Frequency Hopping Cognitive Radio Networks," vol. 32, no. 3, pp. 563–571, 2014.

[37] Y. Tawk, J. Costantine, and C. G. Christodoulou, "Recon fi gurable Filtennas and MIMO in Cognitive Radio Applications," vol. 62, no. 3, pp. 1074–1083, 2014.

[38] Y. Zhao and M. Song, "FMAC : A Fair MAC Protocol for Coexisting Cognitive Radio Networks," pp. 1474–1482, 2013.

[39] A. O. Bicen, S. Member, E. B. Pehlivanoglu, and S. Member, "Dedicated Radio Utilization for Spectrum Handoff and Efficiency in Cognitive Radio Networks," vol. 14, no. 9, pp. 5251–5259, 2015.

[40] M. Jo, L. Han, D. Kim, and H. P. In, "Selfish Attacks and Detection in Cognitive Radio Ad-Hoc Networks," no. June, pp. 46–50, 2013.

[41] N. Of and N. Of, "S ELF -O RGANIZATION P ARADIGMS AND O PTIMIZATION A PPROACHES FOR C OGNITIVE R ADIO T ECHNOLOGIES : A S URVEY U NIVERSITY OF S CIENCE AND T ECHNOLOGY B EIJING," no. April, pp. 36–42, 2013.

[42] C. From, "Open Research Issues in Multi-Hop Cognitive Radio Networks," no. April, pp. 168–176, 2013.

[43] J. Sonnenberg, D. B. Chester, J. Schroeder, and K. Olds, "QUANTIFYING THE RELATIVE MERITS OF GENETIC AND SWARM ALGORITHMS FOR NETWORK OPTIMIZATION IN COGNITIVE RADIO NETWORKS," 2013.

[44] T. Jiang, H. Wang, and A. V Vasilakos, "QoE-Driven Channel Allocation Schemes for Multimedia Transmission of Priority-Based Secondary Users over Cognitive Radio Networks," vol. 30, no. 7, pp. 1215–1224, 2012.

[45] G. Chung, S. Sridharan, S. Vishwanath, S. Member, and C. S. Hwang, "On the Capacity of Overlay Cognitive Radios with Partial Cognition," vol. 58, no. 5, pp. 2935–2949, 2012.

[46] R. Zhou, X. Li, V. Chakravarthy, and Z. Wu, "Software Defined Radio Implementation of SMSE based Overlay Cognitive Radio in High Mobility Environment," no. Ici, 2011.

[47] A. G. Fragkiadakis, E. Z. Tragos, and I. G. Askoxylakis, "A Survey on Security Threats and Detection Techniques in Cognitive Radio Networks," vol. 15, no. 1, pp. 428–445, 2013.

[48] Z. Chen, N. Guo, and R. C. Qiu, "Building A Cognitive Radio Network Testbed," pp. 91–96, 2011.

[49] S. Wang, L. Xie, H. Liu, B. Zhang, and H. Zhao, "ACRA : An Autonomic and Expandable Architecture for Cognitive Radio Nodes," 2010.

[50] M. Khalid and R. Sankar, "Impact of Mobility Prediction on the Performance of Cognitive Radio Networks," 2010.

[51] Y. Zhang, D. Niyato, P. Wang, and E. Hossain, "Auction-based resource allocation in cognitive radio systems," *IEEE Commun. Mag.*, vol. 50, no. 11, pp. 108–120, 2012.

[52] P. Demestichas, A. Georgakopoulos, D. Karvounas, K. Tsagkaris, V. Stavroulaki, and J. Lu, "5G on the Horizon," no. september, pp. 47–53, 2013.

[53] L. Breiman, "No Title," pp. 1–35, 1999.

[54] R. Schapire, "Boosting : Foundations and Algorithms Example : Spam Filtering."

[55] V. Y. Kulkarni, "Random Forest Classifiers : A Survey and Future Research Directions," vol. 36, no. 1, pp. 1144–1153, 2013.

[56] N. Nguyen-Thanh and I. Koo, "A cluster-based selective cooperative spectrum sensing scheme in cognitive radio," *EURASIP J. Wirel. Commun. Netw.*, vol. 2013, no. 1, p. 176, 2013.

[57] R. Singh, A. Kaur, and A. Singh, "Dynamic Cluster based Cooperation for Fair Spectrum Sensing and Sharing in Cognitive Radio Networks," vol. 3, no. 1, pp. 217–225, 2015.

[58] K. Tsagkaris, A. Katidiotis, and P. Demestichas, "Neural network-based learning schemes for cognitive radio systems," vol. 31, pp. 3394–3404, 2008.

[59] T. W. Rondeau, V. Tech, L. V. Tech, C. J. Rieser, J. Hopkins, C. W. Bostian, and V. Tech, "COGNITIVE RADIOS WITH GENETIC ALGORITHMS : INTELLIGENT CONTROL OF SOFTWARE DEFINED RADIOS," 2004.

[60] N. Lavrač, "Hierarchical Clustering of Multiple Decision Trees," no. 1984, pp. 8–11, 2016.

[61] L. Cigler, "Decentralized Anti-coordination Through Multi-agent Learning," vol. 47, pp. 441–473, 2013.

[62] I. W. Communications, "C ognitive R adio S pectrum S ensing F ramework B ased on M ulti -A gent A rchitecture for 5G N etworks," no. June 2016, 2015.

[63] N. Hosey, S. Bergin, and I. Macaluso, "Q-Learning for Cognitive Radios."

[64] "iot-machinelearning." .

[65] J. M. Iii and G. Q. Maguire, "No Title," no. August, pp. 13–18, 1999.

[66] S. Haykin, "Cognitive Radio : Brain-Empowered," vol. 23, no. 2, pp. 201–220, 2005.

[67] Y. Gai, B. Krishnamachari, and R. Jain, "Learning Multiuser Channel Allocations in Cognitive Radio Networks : A Combinatorial Multi-Armed Bandit Formulation," 2010.

[68] M. Pavella and I. Introduction, "Decision tree approach to power systems security assessment," vol. 1, no. 1, 1993.

Text Mining: Techniques, Applications and Issues

Ramzan Talib*, Muhammad Kashif Hanif†, Shaeela Ayesha‡, and Fakeeha Fatima§

Department of Computer Science,
Government College University, Faisalabad, Pakistan

***Abstract*—Rapid progress in digital data acquisition techniques have led to huge volume of data. More than 80 percent of today's data is composed of unstructured or semi-structured data. The discovery of appropriate patterns and trends to analyze the text documents from massive volume of data is a big issue. Text mining is a process of extracting interesting and non-trivial patterns from huge amount of text documents. There exist different techniques and tools to mine the text and discover valuable information for future prediction and decision making process. The selection of right and appropriate text mining technique helps to enhance the speed and decreases the time and effort required to extract valuable information. This paper briefly discuss and analyze the text mining techniques and their applications in diverse fields of life. Moreover, the issues in the field of text mining that affect the accuracy and relevance of results are identified.**

***Keywords*—Classification; Knowledge Discovery; Applications; Information Extraction; Patterns**

I. Introduction

The size of data is increasing at exponential rates day by day. Almost all type of institutions, organizations, and business industries are storing their data electronically. A huge amount of text is flowing over the internet in the form of digital libraries, repositories, and other textual information such as blogs, social media network and e-mails [1]. It is challenging task to determine appropriate patterns and trends to extract valuable knowledge from this large volume of data [2]. Traditional data mining tools are incapable to handle textual data since it requires time and effort to extract information.

Text mining is a process to extract interesting and significant patterns to explore knowledge from textual data sources [3]. Text mining is a multi-disciplinary field based on information retrieval, data mining, machine learning, statistics, and computational linguistics [3]. Figure 1 shows the Venn diagram of text mining and its interaction with other fields. Several text mining techniques like summarization, classification, clustering etc., can be applied to extract knowledge. Text mining deals with natural language text which is stored in semi-structured and unstructured format [4]. Text mining techniques are continuously applied in industry, academia, web applications, internet and other fields [5]. Application areas like search engines, customer relationship management system, filter emails, product suggestion analysis, fraud detection, and social media analytics use text mining for opinion mining, feature extraction, sentiment, predictive, and trend analysis [6].

Generic process of text mining performs the following steps (Figure 2)

- Collecting unstructured data from different sources available in different file formats such as plain text, web pages, pdf files etc.
- Pre-processing and cleansing operations are performed to detect and remove anomalies. Cleansing process make sure to capture the real essence of text available and is performed to remove stop words stemming (process of identifying the root of certain word) and indexing the data [7].
- Processing and controlling operations are applied to audit and further clean the data set by automatic processing.
- Pattern analysis is implemented by Management Information System (MIS).
- Information processed in the above steps are used to extract valuable and relevant information for effective and timely decision making and trend analysis [8].

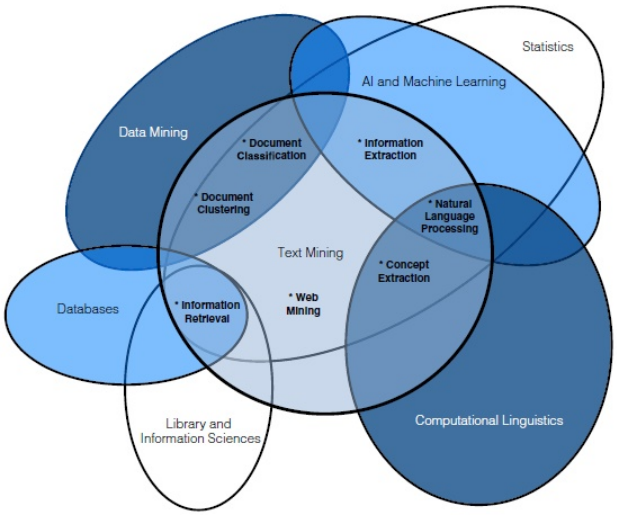


Fig. 1. Venn diagram of text mining interaction with other fields [4]

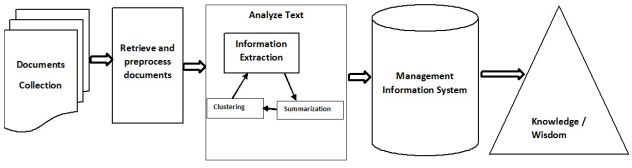


Fig. 2. Text mining process [5]

Extraction of valuable information from a corpus of different document is a tedious and tiresome task. The selection of

appropriate technique for mining text reduce the time and effort to find the relevant patterns for analysis and decision making. The objective of this paper is to analyze different text mining techniques which help to perform text analytics effectively and efficiently from large amount of data. Moreover, the issues that arise during text mining process are identified.

This paper is organized in different sections. Previous work is discussed in section II. In section III, different techniques of text mining are explained. Section IV presents the application areas of text mining techniques. In section V, issues and challenges in text mining field are highlighted. Section VI concludes the outcomes.

II. Review of Literature

[5] described that gathering, extracting, pre-processing, text transformation, feature extraction, pattern selection, and evaluation steps are part of text mining process. In addition, different widely used text mining techniques, i.e., clustering, categorization, decision tree categorization, and their application in diverse fields are surveyed. [8] highlighted the issues in text mining applications and techniques. They discussed that dealing with unstructured text is difficult as compared to structured or tabular data using traditional mining tools and techniques. They have shown the applications of text mining process in bioinformatics, business intelligence and national security system. Natural language processing and entity recognition techniques has reduced the issues that occur during text mining process. However, there exist issues which need attention.

[9] explored MEDLINE biomedical database by integrating a framework for named entity recognition, classification of text, hypothesis generation and testing, relationship and synonym extraction, extract abbreviations. This new framework helps to eliminate unnecessary details and extract valuable information. [10] analyzed the text using text mining patterns and showed term based approaches cannot analyze synonyms and polysemy properly. Moreover, a prototype model was designed for specification of patterns in terms of assigning weight according to their distribution. This approach helps to enhance the efficiency of text mining process. [11] presented a crime detection system using text mining tools and relation discovery algorithm was designed to correlate the term with abbreviation.

[12] presented a top down and bottom up approach for web based text mining process. To combine the similar text documents, they apply k-mean clustering technique for bottom up partitioning. To find out the similarity within the document TF-IDF (Term Frequency- Inverse Document Frequency) algorithm has been used to find information regarding specific subjects. [13] gave an overview of applications, tools and issues arises to mine the text. They discussed that documents may be structured, semi structured or unstructured and extracting useful information is a tiresome task. They presented a generic framework for concept based mining which can be visualized as text refinement and knowledge distillation phases. The intermediate form of entity representation mining depends on specific domain.

[14] presented innovative and efficient pattern discovery techniques. They used the pattern evolving and discovering techniques to enhance the effectiveness of discovering relevant and appropriate information. They performed BM25 and vector support machine based filtering on router corpus volume 1 and text retrieval conference data to estimate the effectiveness of the suggested technique. [15] performed various experiments of classification using multi-word features on the text. They proposed a hand-crafted method to extract multi-word features from the data set. To classify and extract multi-word text they divide text into linear and nonlinear polynomial form in support of vector machine that improve the effectiveness of the extracted data.

III. The Reflective Process

Different text mining techniques are available that are applied for analyzing the text patterns and their mining process [16]. Figure 3 shows the Venn diagram for the inter-relationship among text mining techniques and their core functionality. Document classification (text classification, document standardization), information retrieval (keyword search / querying and indexing), document clustering (phrase clustering), natural language processing (spelling correction, lemmatization, grammatical parsing, and word sense disambiguation), information extraction (relationship extraction / link analysis), and web mining (web link analysis) [6].

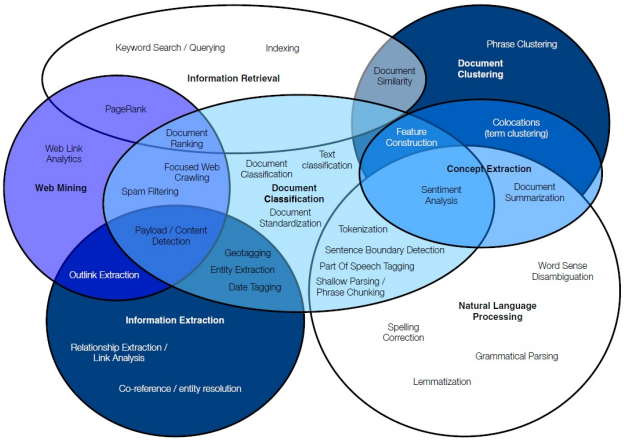


Fig. 3. Inter-relationship among different text mining techniques and their core functionalities [6]

A. Information Extraction

Information Extraction (IE) is a technique that extract meaningful information from large amount of text. Domain experts specify the attributes and relation according to the domain [17]. IE systems are used to extract specific attributes and entities from the document and establish their relationship [18]. The extracted corpus is stored into database for further processing. Precision and recall process is used to check and evaluate the relevance of results on the extracted data. In-depth and complete information about the relevant field is required to perform information extraction process to attain more relevant results [19].

B. Information Retrieval

Information Retrieval (IR) is a process of extracting relevant and associated patterns according to a given set of words

or phrases. There is a close relationship in text mining and information retrieval for textual data. In IR systems, different algorithms are used to track the user's behavior and search relevant data accordingly [19]. Google and Yahoo search engines are using information retrieval system more frequently to extract relevant documents according to a phrase on Web. These search engines use query based algorithms to track the trends and attain more significant results. These search engines provide user more relevant and appropriate information that satisfy them according to their needs [8].

C. Natural Language Processing

Natural language processing (NLP) concerns to the automatic processing and analysis of unstructured textual information. It perform different types of analysis such as Named Entity Recognition (NER) for abbreviation and their synonyms extraction to find the relationships among them [10]. NER identify all the instances of specified object from a group of documents. These entities and their instances allow the identification of relationship and other information to attain their key concept. However, this technique lacks complete dictionary list for all named entities used for identification [9], [10]. Complex query based algorithms need to be used to attain acceptable results. In real world, a single entity has numerous terms like TV and Television. Sometimes, a group of successive words have a multi-word names to identify the boundaries and resolve overlapping issues by using classification technique. Approaches to deal with NER usually fall into four categories: lexicon, rule, statistical based or mixture of these approached. NER systems have achieved the relevance level from 75 to 85 percent [20].

To extract synonym and abbreviation from textual data, co-referencing technique is frequently in use for NLP. Natural Languages (NL) have lot of complexities as a text extracted from different sources don't have identical words or abbreviation. There is a need to detect such issues and make rules for their uniform identification [21]. For example, NER and co-referencing approaches establish a logical relationship to extract and identify the role of person in an organization (use the name of a person at once and then use pronoun instead of name again and again) [22].

D. Clustering

Clustering is an unsupervised process to classify the text documents in groups by applying different clustering algorithms. In a cluster, similar terms or patterns are grouped extracted from various documents. Clustering is performed in top-down and bottom up manner. In NLP, various types of mining tools and techniques are applied for the analysis on unstructured text. Different techniques of clustering are hierarchical, distribution, density, centroid, and k-mean [22].

E. Text Summarization

Text summarization is a process of collecting and producing concise representation of original text documents [23]. Pre-processing and processing operations are performed on the raw text for summarization. Tokenization, stop word removal, and stemming methods are applied for pre-processing. Lexicon lists are generated at processing stage of text summarization. In past, automatic text summarization was performed on the basis of occurrence a certain word or phrase in document. Later on, additional methods of text mining were introduced with standard text mining process to improve the relevance and accuracy of results [11].

To summarize the text documents, weighted heuristics method extract features by following specific rules. Sentence length, fixed phrase, paragraph, thematic word, and upper case word identification features can be implemented and analyzed for text summerization. Text summarization techniques can be applied on multiple documents at the same time. Quality and type of classifiers depend on nature and theme of the text documents [24].

IV. Application of Text Mining

A. Digital Libraries

Numerous text mining techniques and tools are in use to ascertain the patterns and trends from journals and proceedings from immense amount of repositories. These sources of information help in the field of research and development. Libraries are a great source of information for the researchers and digital libraries are endeavoring to the significance of their collection. It provides a novel method of organizing information in such a way that make it possible to available trillions of documents online. It provides a novel way to organize information and make it possible to access millions of documents online. Green-stone international digital library that support multiple languages and multilingual interfaces provide a springy method for extracting documents that handle multiple formats, i.e., Microsoft word, pdf, postscript, HTML, scripting languages and e-mail messages [11]. It also supports the document extraction in the form of audio visual and image format along with text documents. In text mining process various operation are performed like documents selection, enrichment, extracting information and tackling entities among the documents and generating instinctive co-referencing and summarization [25]. GATE, Net Owl and Aylien are frequently used tools for text mining in digital libraries.

B. Academic and Research Field

In education field, various text mining tools and techniques are used to analyze the educational trends in specific region, student's interest in specific field and employment ratio [24]. Use of text mining in research field help to find and classify research papers and relevant material of different fields at one place. The use of k-means clustering and other techniques help to identify the attributes of relevant information. Students performance in different subjects can be accessed and how different attributes effect the selection of subjects [11], [26].

C. Life Science

Life science and health care industries are generating large amount of textual and numerical data regarding patients record, diseases, medicines, symptoms and treatments of diseases and many more. It is a big challenge to filter out an appropriate and relevant text to take a decision from a large biological repository [25]. The medical records contain varying in nature, complex, lengthy and technical vocabulary are used that make the knowledge discovery process very difficult [27]. Text

mining tools in biomedical field provides an opportunity to extract valuable information, their association and inferring relationship among various diseases, species, and genes. Use of an appropriate text mining tools in medical field help to evaluate the effectiveness of medical treatments that show effectiveness by comparing different diseases, symptoms and their course of treatments [28]. Text mining use in biomarker discovery, pharmaceutical industry, clinical trade analysis, pre-clinical safe toxicity studies, patent competitive intelligence and landscaping, mapping of genes diseases and exploring the targeted identifications by using various tools [20].

D. Social Media

Text mining software packages are available for analyzing social media applications to monitor and analyze the online plain text from internet news, blogs, email etc. Text mining tools help to identify and analyze number of posts, likes and followers on the social media network. This kind of analysis show the people reaction on different posts, news and how it spread around. It shows the behavior of people belong to specific age group or communities having similarity and variation in views about the same post [29], [30].

E. Business Intelligence

Text mining plays a significant role in business intelligence that help organizations and enterprises to analyze their customers and competitors to take better decisions. It provides a deeper insight about business and give information how to improve the customer satisfaction and gain competitive advantages [31]. The text mining tools like IBM text analytics, Rapid miner, GATE help to take decisions about the organization that generate alerts about good and bad performance, market changeover that help to take remedial actions. It also helps in telecommunication industry, business and commerce applications and customer chain management system [32].

V. Issues in Text Mining Field

Many issues occur during the text mining process and effect the efficiency and effectiveness of decision making. Complexities can arise at the intermediate stage of text mining. In pre-processing stage various rules and regulations are defined to standardize the text that make text mining process efficient. Before applying pattern analysis on the document there is a need to convert unstructured data into intermediate form but at this stage mining process has its own complications. Sometime real theme or data mislay its importance due to the modification in the text sequence [27]. Another major issue is a multilingual text refinement dependency that create problems. Only few tools are available that support multiple languages [33]. Various algorithms and techniques are used independently to support multilingual text. Because numerous important documents persist outside the text mining process because various tools dont support them. These issues create a lots of problems in knowledge discovery and decision making process. Infect real benefit is difficult to attain by using the existing text mining techniques and tools because its rarely support multilingual documents [34].

Integration of domain knowledge is an important area as it performs specific operations on specified corpus and attain desired outcomes. In this situations domain knowledge from which document corpus to be extracted need to integrate with the computing abilities from which information have to be attained. According to the requirements of the field, experts are needed to work collaboratively from diverse domains to extract more effective, precise and accurate results [22], [27].

The use of synonyms, polysems and antonyms in the documents create problems (abstruseness) for the text mining tools that take both in the same context. It is difficult to categorize the documents when collection of document is large and generated from diverse fields having the same domain. Abbreviations gives changed meaning in different situation is also a big issue [35]. Varying concepts of granularity change the context of text according to the condition and domain knowledge. There is need to describe rules according to the field that will be used as a standard in the area and can be embedded in text mining tools as a plug-in. It entails lots of effort and time to develop and deploy plug-ins in all fields separately. To develop plug-ins in depth and proper knowledge about the specific domain will be required [34], [36]. Natural languages have lots of complications in itself that create problem in text refinement methods and the identification of entity relationship. Words having same spelling but give diverse meaning, for example, fly and fly. Text mining tools considered both as similar while one is verb and other is noun. Grammatical rules according to the nature and context is still an open issue in the field of text mining [36].

VI. Conclusion

The availability of huge volume of text based data need to be examined to extract valuable information. Text mining techniques are used to analyze the interesting and relevant information effectively and efficiently from large amount of unstructured data. This paper presents a brief overview of text mining techniques that help to improve the text mining process. Specific patterns and sequences are applied in order to extract useful information by eliminating irrelevant details for predictive analysis. Selection and use of right techniques and tools according to the domain help to make the text mining process easy and efficient. Domain knowledge integration, varying concepts granularity, multilingual text refinement, and natural language processing ambiguity are major issues and challenges that arise during text mining process. In future research work, we will focus to design algorithms which will help to resolve issues presented in this work.

References

[1] R. Sagayam, A survey of text mining: Retrieval, extraction and indexing techniques, *International Journal of Computational Engineering Research*, vol. 2, no. 5, 2012.

[2] N. Padhy, D. Mishra, R. Panigrahi *et al.*, "The survey of data mining applications and feature scope," *arXiv preprint arXiv:1211.5723*, 2012.

[3] W. Fan, L. Wallace, S. Rich, and Z. Zhang, "Tapping the power of text mining," *Communications of the ACM*, vol. 49, no. 9, pp. 76–82, 2006.

[4] S. M. Weiss, N. Indurkhya, T. Zhang, and F. Damerau, *Text mining: predictive methods for analyzing unstructured information*. Springer Science and Business Media, 2010.

[5] S.-H. Liao, P.-H. Chu, and P.-Y. Hsiao, "Data mining techniques and applications–a decade review from 2000 to 2011," *Expert Systems with Applications*, vol. 39, no. 12, pp. 11 303–11 311, 2012.

[6] W. He, "Examining students online interaction in a live video streaming environment using data mining and text mining," *Computers in Human Behavior*, vol. 29, no. 1, pp. 90–102, 2013.

[7] G. King, P. Lam, and M. Roberts, "Computer-assisted keyword and document set discovery from unstructured text," *Copy at http://j.mp/1qdVqhx Download Citation BibTex Tagged XML Download Paper*, vol. 456, 2014.

[8] N. Zhong, Y. Li, and S.-T. Wu, "Effective pattern discovery for text mining," *IEEE transactions on knowledge and data engineering*, vol. 24, no. 1, pp. 30–44, 2012.

[9] A. Henriksson, H. Moen, M. Skeppstedt, V. Daudaravičius, and M. Duneld, "Synonym extraction and abbreviation expansion with ensembles of semantic spaces," *Journal of biomedical semantics*, vol. 5, no. 1, p. 1, 2014.

[10] B. Laxman and D. Sujatha, "Improved method for pattern discovery in text mining," *International Journal of Research in Engineering and Technology*, vol. 2, no. 1, pp. 2321–2328, 2013.

[11] C. P. Chen and C.-Y. Zhang, "Data-intensive applications, challenges, techniques and technologies: A survey on big data," *Information Sciences*, vol. 275, pp. 314–347, 2014.

[12] R. Rajendra and V. Saransh, "A Novel Modified Apriori Approach for Web Document Clustering," *International Journal of Computer Applications*, pp. 159–171, 2013.

[13] K. Sumathy and M. Chidambaram, "Text mining: Concepts, applications, tools and issues-an overview," *International Journal of Computer Applications*, vol. 80, no. 4, 2013.

[14] P. J. Joby and J. Korra, "Accessing accurate documents by mining auxiliary document information," in *Advances in Computing and Communication Engineering (ICACCE), 2015 Second International Conference on*. IEEE, 2015, pp. 634–638.

[15] Z. Wen, T. Yoshida, and X. Tang, "A study with multi-word feature with text classification," in *Proceedings of the 51st Annual Meeting of the ISSS-2007, Tokyo, Japan*, vol. 51, 2007, p. 45.

[16] V. Gupta and G. S. Lehal, "A survey of text mining techniques and applications," *Journal of emerging technologies in web intelligence*, vol. 1, no. 1, pp. 60–76, 2009.

[17] R. Agrawal and M. Batra, "A detailed study on text mining techniques," *International Journal of Soft Computing and Engineering (IJSCE) ISSN*, pp. 2231–2307, 2013.

[18] D. S. Dang and P. H. Ahmad, "A review of text mining techniques associated with various application areas," *International Journal of Science and Research (IJSR)*, vol. 4, no. 2, pp. 2461–2466, 2015.

[19] R. Steinberger, "A survey of methods to ease the development of highly multilingual text mining applications," *Language Resources and Evaluation*, vol. 46, no. 2, pp. 155–176, 2012.

[20] A. M. Cohen and W. R. Hersh, "A survey of current work in biomedical text mining," *Briefings in bioinformatics*, vol. 6, no. 1, pp. 57–71, 2005.

[21] E. A. Calvillo, A. Padilla, J. Muñoz, J. Ponce, and J. T. Fernandez, "Searching research papers using clustering and text mining," in *Electronics, Communications and Computing (CONIELECOMP), 2013 International Conference on*. IEEE, 2013, pp. 78–81.

[22] B. L. Narayana and S. P. Kumar, "A new clustering technique on text in sentence for text mining," *IJSEAT*, vol. 3, no. 3, pp. 69–71, 2015.

[23] B. A. Mukhedkar, D. Sakhare, and R. Kumar, "Pragmatic analysis based document summarization," *International Journal of Computer Science and Information Security*, vol. 14, no. 4, p. 145, 2016.

[24] R. Al-Hashemi, "Text summarization extraction system (tses) using extracted keywords." *Int. Arab J. e-Technol.*, vol. 1, no. 4, pp. 164–168, 2010.

[25] I. H. Witten, K. J. Don, M. Dewsnip, and V. Tablan, "Text mining in a digital library," *International Journal on Digital Libraries*, vol. 4, no. 1, pp. 56–59, 2004.

[26] S. Ayesha, T. Mustafa, A. R. Sattar, and M. I. Khan, "Data mining model for higher education system," *Europen Journal of Scientific Research*, vol. 43, no. 1, pp. 24–29, 2010.

[27] A. Henriksson, J. Zhao, H. Dalianis, and H. Boström, "Ensembles of randomized trees using diverse distributed representations of clinical events," *BMC Medical Informatics and Decision Making*, vol. 16, no. 2, p. 69, 2016.

[28] I. Alonso and D. Contreras, "Evaluation of semantic similarity metrics applied to the automatic retrieval of medical documents: An umls approach," *Expert Systems with Applications*, vol. 44, pp. 386–399, 2016.

[29] C. Ding and H. Peng, "Minimum redundancy feature selection from microarray gene expression data," *Journal of bioinformatics and computational biology*, vol. 3, no. 02, pp. 185–205, 2005.

[30] Y. Zhao, "Analysing twitter data with text mining and social network analysis," in *Proceedings of the 11th Australasian Data Mining and Analytics Conference (AusDM 2013)*, 2013, p. 23.

[31] F. Fatima, Z. W. Islam, F. Zafar, and S. Ayesha, "Impact and usage of internet in education in pakistan," *European Journal of Scientific Research*, vol. 47, no. 2, pp. 256–264, 2010.

[32] R. Sharda and M. Henry, "Information extraction from interviews to obtain tacit knowledge: A text mining application," *AMCIS 2009 Proceedings*, p. 283, 2009.

[33] H. Solanki, "Comparative study of data mining tools and analysis with unified data mining theory," *International Journal of Computer Applications*, vol. 75, no. 16, 2013.

[34] A. Kumaran, R. Makin, V. Pattisapu, and S. E. Sharif, "Automatic extraction of synonymy information:-extended abstract," *OTT06*, vol. 1, p. 55, 2007.

[35] A. Kaklauskas, M. Seniut, D. Amaratunga, I. Lill, A. Safonov, N. Vatin, J. Cerkauskas, I. Jackute, A. Kuzminske, and L. Peciure, "Text analytics for android project," *Procedia Economics and Finance*, vol. 18, pp. 610–617, 2014.

[36] N. Samsudin, M. Puteh, A. R. Hamdan, and M. Z. A. Nazri, "Immune based feature selection for opinion mining," in *Proceedings of the World Congress on Engineering*, vol. 3, 2013, pp. 3–5.

A Generic Model for Assessing Multilevel Security-Critical Object-Oriented Programs

Bandar M. Alshammari
Department of Information Technology
School of Computer and Information Sciences
Aljouf University, Saudi Arabia

***Abstract*—The most promising approach for developing secure systems is the one which allows software developers to assess and compare the relative security of their programs based on their designs. Thereby, software metrics provide an easy approach for evaluating the security of certain object-oriented designs. They can also measure the impact on security that caused by modifications to existing programs. However, most studies in this area focus on a binary classification of data, either is classified or unclassified. In fact, there are other models with other classifications of data, for instance, the common model used by Defense departments that classifies data into four security levels. However, these various classifications have received little attention in terms of measuring their effect. This paper introduces a model for measuring information flow of security-critical data within a certain object-oriented program with multilevel classification of its security-critical data. It defines a set of object-oriented security metrics which are capable of assessing the security of a given program's design from the point of view of potential information flow. These metrics can be used to compare the security of programs or assess the effect of program modifications on security. Specifically, this paper proposes a generic model that consists of several security metrics to measure the relative security of object-oriented designs with respect to design quality properties of accessibility, cohesion, coupling, and design size.**

***Keywords*—Multilevel Security Models; Object-Orientation; Security Metrics; Security Matrix; Unified Model Language**

I. Introduction

Recently, security has become one of the most crucial aspects of systems' development due to the increasing number of risks and breaches which systems are facing. Therefore, several studies have suggested different approaches for reducing these risks and preventing the existence of vulnerabilities. One example defines a number of safe coding practices which a programmer needs to follow in order to have a more secure product [1] [2]. Another approach relies on previously defined security vulnerabilities. This approach statically analyses programs' codes to check if the they contain any of the defined vulnerabilities [3] [4] [5] [6].

Another approach which can be easily applied is to define security metrics that can quantify the security level of certain programs. [7]. Security metrics can be an effective tool in helping software programers to identify level of risks in a given program. They also provide programmers with guidance of how to raise awareness within the organisation. Security metrics that are based on the designs of programs can provide systems' developers with an early guidance for discovering vulnerabilities and guiding them on following certain secure corrective steps.

Most of the work on security metrics focus on classifying data into two levels, either security-critical or not. However, this is not the case in real systems. Real security-critical systems have multiple level of classifications for their sensitive data. Therefore, this paper defines a generic model that takes into consideration this aspect. It studies the impact on security of four of the most common software design properties, which are used in order to enhance the software quality. These properties consist of Data Encapsulation, Cohesion, Coupling, and Design Size [8]. In this work, a number of security metrics for object-oriented designs are defined with respect to those quality design properties. These metrics are capable of quantifying the security level of certain programs with regard to the potential flow of security-critical information based on the security design principles of "reducing the size of the attack surface" [9] [10] [11] and " least privilege" [12] [2]. Such metrics will also assist software developers in assessing the impact on security of any modifications occurred to their programs.

The remainder of this paper is organised as follows. Section 2 shows the related work and how this paper defines a model that is distinct from previous work. Section 3 shows the research methodology that is used to define this model. Section 4 explains the assumptions which need to be considered when applying this model. Section 5 studies the relevant security design principles that this model considers. Section 6 illustrates the model defined by this paper and the defined security design metrics. Section 7 illustrates a case study of how to apply this model to a real system based on its design, and it also shows the results of the security metrics. Finally, Section 8 concludes the paper and explains future extensions of this work.

II. Related Work

Many researchers have proposed different approaches which help in developing more secure programs. One of these approaches defines a list of principles that can be used as guidance for developing secure systems [12] [2]. Other approaches have developed general coding principles that aim to discover code vulnerabilities [1] [13]. However, these approaches cannot discover security risks at an early stage and are not capable of quantifying security of given programs.

The approach which sounds promising with this regard is to define security metrics at an early stage of development. Some of the work in this area has proposed metrics for measuring software security of object-oriented programs such as the work of [14] [15] [16] [17] [18] [19] [20]. The main objective of these metrics is to assess programs' security at different stages

of its development life cycle in order to give a prediction of the existing vulnerabilities in the program.

Multilevel security, on the other hand, is a crucial aspect to information systems. In fact, many admit that it is a must for any system to have various levels of security. In other words, different security-critical components of any system have to be classified into different levels of criticality in terms of information security. There exists several security models for classifying information with regard to their security sensitivity. For instance, the US model defines four level of security classifications (Top Secret, Secret, Confidential, and unclassified)[1]. Another example is the British which uses a model of six levels of security classifications (Top Secret, Secret, Confidential, Restricted, Protected, and unclassified)[2].

Unfortunately, most of studies on software security metrics focus on defining metrics by classifying system's components to classified and not classified ones. Furthermore, there exists some studies that study multilevel security such as the work of Kotenko and Doynikova [21] which defines security metrics for various levels of the system. However, this work doesn't consider the information flow of various security values of security-critical data. Instead, it focuses on dividing different types of metrics for different parts of the system [21].

None of existing projects have developed metrics for program security based on its design artifacts that takes into consideration the different security level of the system's components. This paper defines a generic model that is applicable to any security-critical object-oriented program in order to assess its security with respect to the potential flow of security-critical data.

III. Research Methodology

The approach of Ourston and Monney's [22] states that any project's implementation should be conducted in two steps. The first one is the analytical part which is about defining the theory of the project. The second one is called the empirical part which is conducted to prove the analytical part. This paper follows this approach, and hence it defines the first part that is called *what to measure* and the second part which is called *how to measure*.

A. What to measure

This paper aims to define a model that is capable of assessing the security of object-oriented programs with multilevel of data security classifications. It concentrates on specific object-oriented design properties which have the most effect on security of programs. The proposed metrics need to measure the security of programs by identifying the potential information flow of security-critical data.

B. How to measure

This paper defines a number of security metrics for object-oriented programs with multilevel security classifications of their components. The developed metrics have to adhere to certain security design principles for developing secure programs. These metrics can be applied to any object-oriented program as long as its class diagram is provided. Such metrics will assist software developers in assessing the security level of their programs from an early stage of development based on the design artifacts of the programs.

Table I. Model Terminology

Name	Description
Security-Critical Attribute	An attribute which holds confidential data.
Security-Critical Method	A function which accesses or interacts with security-critical attributes.
Security-Critical Class	A class which has at least one security-critical attribute or one security-critical method.

IV. Model Assumptions

The model defined in this paper aims to introduce a set of security metrics for programs with multilevel classifications of data secrecy based on their designs. These metrics measure the potential flow of security-critical data of a given object-oriented design. Each metric is defined in relation with a specific security design principle that needs to adhere to in order to achieve a secure program. The model focuses on defining such metrics with respect to four object-oriented properties (i.e., data encapsulation, cohesion, coupling, and design size) [8]. These metrics are a comparative measurement which means that results of these metrics can be used to compare the relative security of various object-oriented programs with regard to these four design properties.

The metrics have been designed so that their results are within the range 0 to 1. As the value of the metric decreases, the more secure a program is. Similarly, as the value of the metric increases, the less secure a program is. This means that lowers values of these metrics, the more they adhere to their related security design principles. And higher values mean that they less adhere to their relevant security design principles.

One approach which can be used to apply security metrics based on the program's design is developed by Alshammari et al. [19] [20]. This approach depends on accurately providing annotated classes of security-critical data for a given object-oriented program using UMLsec [23] and SPARKS [24] annotations. In this approach, UMLsec annotations are used to annotate attributes, methods, and classes with *"'«secrecy»"'* if they contain or interact with security-critical data. SPARKS annotations, moreover, are used to show the interactions of methods with security-critical attributes, methods, and classes.

Instead, this model develops a new approach which helps in applying security design metrics defined here. This approach relies on the program's developers to provide an accurately designed two dimensional matrix. This matrix shows the security level of all interactions with security-critical classes that are caused by methods or classes in a given design.

To illustrate this approach, a class is extracted from a certain program called Student Information System is shown in Figure 1 and the matrix related to this class is shown in Figure 2. Suppose the security model used in this context is $\{0,1,2,3\}$; with 0 being not a security-critical attribute, and 3 is the most security-critical attribute. In this class,

[1] http://www.state.gov/m/ds/clearances/c10977.htm#5

[2] https://www.gov.uk/government/publications/government-security-classifications

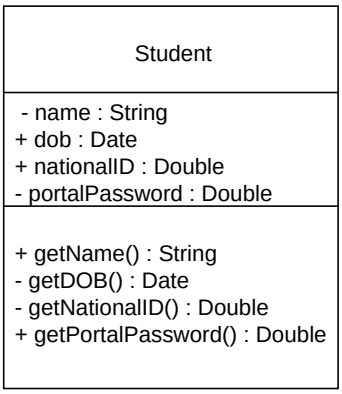


Figure 1. Student UML Class Diagram

		Attributes			
		- name	+ dob	+ nationalID	- portalPassword
Methods	+ getName()	0	0	0	0
	- getDOB()	0	1	0	0
	- getNationalID()	0	0	2	0
	+ getPortalPassword()	0	0	0	3

Figure 2. Student Class Security Matrix

suppose that `name` is not a security-critical attribute (hence, its security value is 0), and is accessed by method `getName`. `DOB` is supposed to be a security-critical attribute with security value 1, and is only accessed by method `getDOB`. Also suppose that `nationalID` is another security-critical attribute with security value 2, and is only accessed by method `getNationalID`. Suppose that `portalPassword` is also a security-critical attribute with security value 3, and is only accessed by method `getPortalPassword`. Then, the security matrix associated with the class diagram from Figure 1can be seen in Figure 2.

V. Relevant Security Design Principles

This section illustrates those security design principles that are relevant to the security design metrics defined by this model. In this paper, two design principles have been chosen to be studied when developing this model (i.e., Least Privilege and Reduce Attack surface). These were chosen based on previous work [19] which has shown that these two principles can have the most effect on developing secure systems, and hence they need to be intensively considered when defining security metrics. Therefore, these metrics are constructed with respect to these two principles in order to measure the security of designs from the perspective of potential information flow.

A. Least Privilege

This principle is described as allowing programs and users to complete a certain job with the least privileges [12] [2]. Its main advantages consist of minimizing interactions between privileged components in a given system, and hence minimizing loss in case of a successful attack [12] [2].

B. Reduce Attack surface

This principle also aims to make programs more secure by decreasing the number of components that can be reached from outside the system [9] [25]. There are several approaches for reducing such components. A common approach is described by Howard [9] suggest reducing the amount of running code by turning off any unnecessary features of the system. Another approach is to minimise the number of entry points in the system that can be accessed by untrusted users [9].

VI. Multilevel Security Assessment Model

Due to the importance of having multiple levels of data security for any systems, this paper defines a model that is capable of defining a set of security metrics that meet this purpose. This model considers four of the most common software design properties to define these metrics. these properties include the data encapsulation, cohesion, coupling, and design size of systems [8]. These metrics depend on the security levels defined by system's security analysts. Furthermore, the model develops a security matrix that can be used easily to extract the required information of these security design metrics.

These security design metrics are defined to be generic, which means that they can be applied to any program regardless of the type of security model it uses. For the purpose of illustration, let's address the set of values of the security model used in this context as $V = \{0, \ldots, v\}$, where an attribute labelled as 0 means it is not security-critical and an attribute labelled as v (the maximum number in the model) means it is the most security-critical attribute in that program. (let the magnitude operator $|S|$ returns the size of a given set S.) These metrics are defined as follows.

A. Accessibility of Security-Critical Attributes Metric (ASCAM)

The design property of data encapsulation in object-oriented designs is concerned with having restrictions on data accessibility inside a given program [8]. In terms of security, it has been shown that this property has a major effect on program's security including the work of Maruyama et al. [13] and Alshammari et al. [19]. These studies have shown that creating security-critical attributes less accessible from outside their classes make programs more secure. This will eventually satisfy the specifications of the security principle of "*reducing the attack surface size*".

This metric aims to measure the proportion of security-critical attributes which are accessible from outside their class in a given design for an object-oriented program with respect to their security levels. Therefore, this metric is defined as; "*The ratio of the number of criticality of non-private security-critical attributes in a given design to the total number of criticality of security-critical attributes in that design*".

Consider the set of attributes in a design D as A_i, $i \in \{1, \ldots, a\}$, set of security-critical attributes in D as SCA_j, $j \in \{1, \ldots, sca\}$ such that $SCA \subseteq A$, and set of non-private

(accessible) security-critical attributes in D as $ASCA_k$, $k \in \{1,\ldots,asca\}$ such that $ASCA \subseteq SCA \subseteq A$. Then, *ASCAM* is expressed as:

$$ASCAM(D) = \frac{\sum_{k=1}^{asca}(ASCA_k \times v)}{|SCA| \times \max(V)} \quad (1)$$

B. Accessibility of Security-Critical Methods Metric (ASCMM)

As it's shown previously that the design property of data encapsulation in object-oriented designs is concerned with having restrictions on data accessibility inside a given program [8]. Another possible way of accessing data is through methods which have access to attributes. Similar to direct accessibility of security-critical attributes, accessing security-critical attributes indirectly through methods which interact with them can have the same security impact. Therefore, it is recommended to have less accessibility to methods which interact with security-critical attributes in order to satisfy the security design principle of "*reducing the attack surface size*" [19].

This metric aims to measure the proportion of methods which have access or interaction with security-critical attributes and are accessible from outside their class in a given object-oriented design with respect to their security levels. Therefore, this metric is defined as; "*The ratio of the number of criticality of non-private security-critical methods in a given design to the total number of criticality of security-critical methods in that design*".

Consider the set of methods in a design D as M_i, $i \in \{1,\ldots,m\}$, set of security-critical methods in D as SCM_j, $j \in \{1,\ldots,scm\}$ such that $SCM \subseteq M$, and set of non-private (accessible) security-critical methods in D as $ASCM_k$, $k \in \{1,\ldots,ascm\}$ such that $ASCM \subseteq SCM \subseteq M$ (Given a set S, let the magnitude operator $|S|$ returns the size of the set.) Then, *ASCMM* is expressed as:

$$ASCMM(D) = \frac{\sum_{k=1}^{ascm}(ASCM_k \times v)}{|SCM| \times \max(V)} \quad (2)$$

C. Cohesiveness of Security-Critical Methods Metric (CSCMM)

In terms of object-oriented design properties, it is recommended to have a cohesive design which increases the interactions of attributes with methods inside their class. In regard to security, it is recommended to decrease cohesiveness of interactions between security-critical attributes and methods inside their classes in order to have more secure programs [19]. Cohesiveness is about privileges over security-critical attributes and the fewer number of such interactions, the more this adheres to the security design principle of least privilege [19].

The main aim of this metric is to measures the degree of potential flow of security-critical data caused by interactions between security-critical attributes and methods in a given object-oriented design taking into consideration the security levels of these attributes. Therefore, this metric is defined as; "*The ratio of the number of methods which interact with security-critical attributes for every class to the maximum number of methods which could interact with attributes for every class in a specific program's design*".

To calculate this metric, the number of interactions with every security-critical attribute is multiplied by its security level for every class. Then, the sum of these values are divided by the total number of possible ways of interactions with attributes in that class. This can be calculated by multiplying the total number of methods in that class by the total number of attributes in the same class. Then, it is multiplied by the maximum value of the security model used in context. The values of all of the classes' cohesiveness is summed, and then divided by the number of classes in the design in order to take the design's average cohesiveness.

Consider the set of classes in a design D as C_i, $i \in \{1,\ldots,c\}$, set of methods in the same design as M_j, $j \in \{1,\ldots,m\}$, set of attributes in D design as A_k, $k \in \{1,\ldots,a\}$, and set of security-critical attributes in D as SCA_l, $l \in \{1,\ldots,sca\}$ such that $SCA \subseteq A$. Let $\alpha(SCA_l)$ be the number of methods which access security-critical attribute SCA_l. Then, *CSCMM* is expressed as:

$$CSCMM(D) = \frac{\sum_{i=1}^{c} \frac{\sum_{l=1}^{sca} \alpha(SCA_l) \times v}{|M| \times |SCA| \times \max(V)}}{|C|} \quad (3)$$

D. Coupling of Security-Critical Classes Metric (CSCCM)

Coupling is one of the most common object-oriented design properties that have been widely examined in several studies. This property is defined by the number of interactions an object has with others inside the program [26]. Many studies have shown that it's recommended for object-oriented programs to be loosely-coupled between its components in order to be more reusable, understandable, and extensible [8] [26].

With regard to information security, it has been shown that there is a high correlation between coupling and insecurity of programs. The study of Liu and Traore [27] has shown that successful attacks in many cases are caused by highly-coupled objects. Furthermore, the studies of Alshammari et al. [19] [20] have shown that loosely-coupled programs can decrease the potential flow of security-critical information, and thus creating more secure programs. This satisfies the security design principle of "*least privilege*" [12].

Therefore, this model focuses on the effect of coupling on programs security, and it defines a security metric that fits the multilevel security model described here. This coupling metric aims to measure the potential occurrence of links between security-critical attributes and classes in a given design taking into consideration the security level of each link. Therefore, this metric is defined as; "*The ratio of criticality of the number of associations of all classes with security-critical attributes to the maximum number of associations which could occur with security-critical attributes for every class in a specific program's design*".

Consider the set of classes in a design D as C_i, $i \in \{1,\ldots,c\}$, set of attributes in the design D as A_k, $k \in \{1,\ldots,a\}$, and set of security-critical attributes in D as SCA_l, $l \in \{1,\ldots,sca\}$ such that $SCA \subseteq A$. Let $\beta(SCA_l)$ be the number of classes which are associated with security-critical attribute SCA_l. Then, *CSCCM* is expressed as:

$$CSCCM(D) = \frac{\sum_{l=1}^{sca} \beta\left(SCA_l\right) \times v}{|C-1| \times |SCA| \times \max(V)} \quad (4)$$

E. Security-Critical Design Size Metric (SCDSM)

Design size is one of the object-oriented properties that must be considered from an early stage of a program's development life-cycle since it has a major impact on its reusability and functionality [8]. Due to this importance, Bansiya and Davis [8] defined a metric that is related to this property. This metric is called Design Size in Classes (DSC) which measures of the number of classes in a specific design [8].

In terms of programs' security, the effect of the design size property has been studied in a number of studies. This include the work of Chowdhury et al. [7] which defined a metric to measure the ratio of critical code segments in a particular program's code. The work of Alshammari et al. [19] [20] have also studied this property. They show that it is desirable to have a small design size of security-critical classes in order to adhere to the security design principle of "*reducing the attack surface size*" [20].

The model defined in this paper considers the importance of the design size property on security, and hence it defines a metric for this purpose. The main goal of this metric is to measure the proportion of security-critical classes taking into consideration their security levels. Therefore, this metric is defined as; "*The ratio of criticality of the number of security-critical classes to the total number of security-critical classes in a specific program's design*".

Consider the set of classes in a design D as C_i, $i \in \{1,\ldots,c\}$, set of attributes in the design D as A_k, $k \in \{1,\ldots,a\}$, set of security-critical attributes in D as SCA_l, $l \in \{1,\ldots,sca\}$ such that $SCA \subseteq A$, and set of classes with security-critical classes defined in them in the design D as SCC_l, $l \in \{1,\ldots,scc\}$ such that $SCC \subseteq C$. Then, $SCDS$ is defined as:

$$SCDSM\left(D\right) = \frac{\sum_{l=1}^{scc} \left(SCC_l \times v\right)}{|C| \times \max(v)} \quad (5)$$

VII. Model Case Study

This section shows how to apply the multilevel security design metrics defined here for a specific banking system. To measure the security of a given program with regard to these metrics, a complete UML class diagram needs to be provided by the system's developers. It is also required to provide the security matrix defined by this model as explained in the previous section. This matrix shows the interactions of every attribute with other methods and classes, and the security level of each interaction.

A. Banking System UML Class Diagram

The UML class diagram shown in Figure 3 is part of a banking system extracted to show how to measure the security of a certain program with regard to the model defined by this paper. This diagram consists of classes that are responsible for storing information about clients of a certain bank. The class diagram consists of five classes; Customers, Credit Cards, Debit Cards, Loans, and Cheques. Each of these classes is responsible for a specific task, and it contains a number of attributes and methods to surf this purpose.

For instance, the Customers class is responsible for storing personal information about clients such as their name, date of birth, national ID, and ebanking password. It also defines a number of methods that are related to retrieve specific data in response to requests from either inside or outside the class. The Loans class stores information about the customer's loans such as the loan amount. It has also functions that are responsible for returning this information once requested. The Cheques class provides information about the customer's cheques, and defines functions that access such information. The Credit Card and Debit Card classes define attributes that store cards' numbers and their passwords. They also contain methods that return and update this information.

B. Banking System Matrix

This section explains how to construct the security matrix associated with the banking system illustrated in this paper. This matrix shows the interactions between attributes and methods in the entire system. It also shows those classes which directly access certain attributes from outside their classes. This security matrix has to include a security value for every interaction. This value is the same as the security level of the attribute which the interaction has occurred with.

For the purpose of this case study, let's suppose that the security model used by in this context is $\{0,1,2,3,4\}$. Attributes which their security value is 0 indicate that they are not security-critical. Other than that, attributes are security-critical and their criticality level depend on their security value. This means as this value increases, the security criticality of a certain attribute increases. For instance, attributes assigned the security value of 1 indicate that they have the least security-critical data, while attributes which their security value is 4 show the highest security-critical data. Moreover, the matrix shows the accessibility of every attribute, method, and class in the system. This information is extracted from access modifiers shown in the UML class diagram of the system.

Table II. Banking System's Attributes Security Levels

Class	Attribute	Security Level
Customers	name	0
Customers	dob	2
Customers	nationalID	3
Customers	portalPassword	4
Credit Cards	cardNo	4
Credit Cards	cardPassword	4
Debit Cards	cardNo	3
Debit Cards	cardPassword	4
Cheques	chequeNo	1
Cheques	chequeAmount	2
Loans	interestRate	0
Loans	loanAmount	1

In the UML class diagram in Figure 3, there are five classes and each one of them has a number of attributes with different security levels. Let's suppose that the system's analyst has provided Table II which shows the security levels for all attributes in the system. These values are used in order to construct the security matrix of the banking system as shown in Figure 4 (which needs to be constructed by the system's analyst).

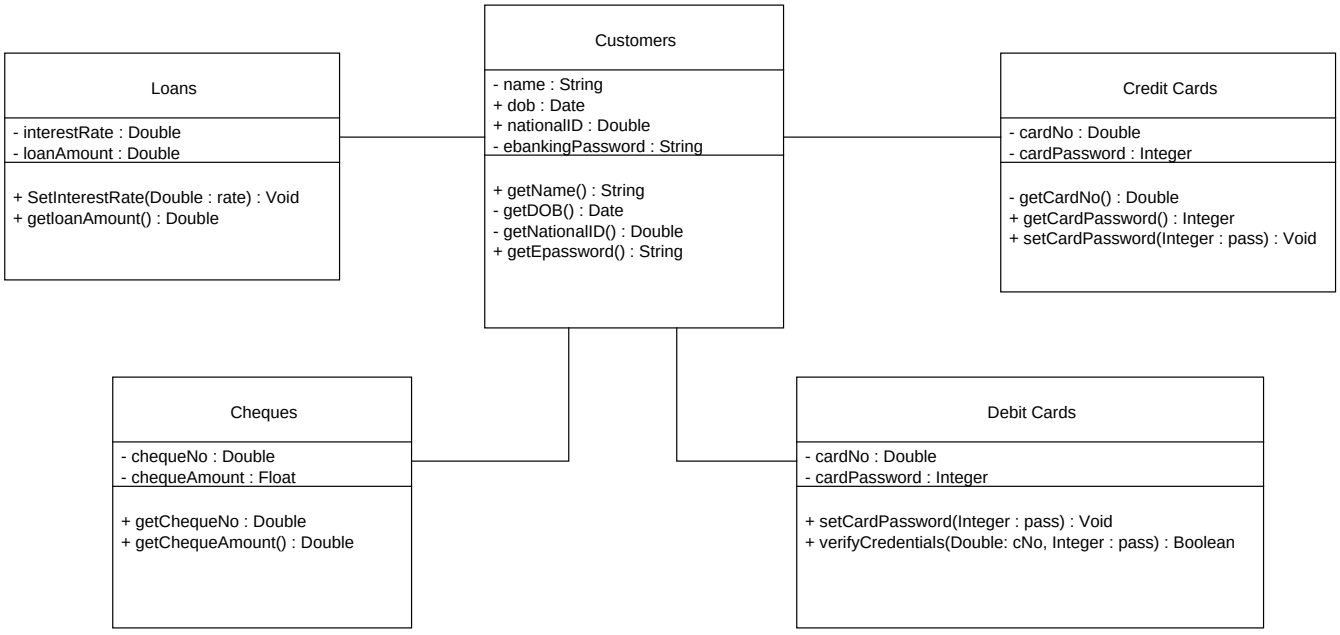


Figure 3. Banking System UML Class Diagram

C. Model Security Metrics Results

The main goal of this section is to show how to calculate the security metrics defined by this model based on the security matrix shown in Figure 4. The purpose and result of each metric are explained here in order to identify their impact on security of the banking system examined here.

The data encapsulation metrics measure the ratio of accessible security-critical attributes and methods in a given class with regard to the total number of security-critical attributes and methods in that class taking into consideration their security values. This means that when calculating these metrics, the security level of each attribute and method has been considered. These values can be easily obtained from the security matrix shown in Figure 4. Therefore, the ASCAM is computed by counting the number of non-private security-critical attributes multiplied by each of their security values. In Design 3, there are eight security-critical attributes out of ten attributes defined in the banking system UML class diagram. Out of these eight security-critical attributes, there are two attributes which are accessible from outside their classes (i.e., `Customers.dob` and `Customers.nationalID`). This information is shown in the security matrix since these two attributes have the (+) symbol in front of them which is an access modifier for public attributes. The security value associated with `Customers.dob` is 2 and with `Customers.nationalID` is 3. Therefore,, the total security value of accessible security-critical attributes in this case is 5. The the total security value of possible accessibility of security-critical attributes is counted by multiplying the total number of security-critical attributes in the design (i.e., 8) to the maximum value of criticality in the security model used in context (i.e., 4) which is 32. Hence, the ASCAM is computed by dividing 5 to 32, which is 0.16.

With regard to the ASCMM which measures the accessibility of security-critical methods in a certain design, it is computed by counting the number of non-private security-critical methods multiplied by their security values for each of them to the total number of possible accessibility of security-critical methods in the same design multiplied by the maximum security value defined in the model. The security value of a certain method is the same as the security value of the attribute it interacts with, which is shown in the security matrix. If there is a method that interacts with more than one attribute of different security levels, then the security level of this method is the same as the highest attribute that it interacts with. The security matrix in Figure 4 shows that there are eleven security-critical methods which eight of them are non-private security-critical methods. Out of these eight methods, there are two methods which their security value is 1; `Cheques.getChequeNo` and `Loans.getLoanAmount`. There is one method which its security value is 2 (i.e., `Cheques.getChequeAmount`). There is one method which its security value is 3; `Customers.getNationalID`. There are five methods which their security value is 4; `Customers.getPortalPassword`, two methods in Credit Cards class `CreditCards.getCardPassword` and `CreditCards.setCardPassword`, and two methods in Debit Cards class`DebitCards.setCardPassword` and `DebitCards.verifyCredentials`. The total possible security value of accessible security-critical methods is counted by multiplying the total number of security-critical methods (i.e., 11) to the maximum value of criticality in the security model used in context (i.e., 4), which is 44. Hence, the ASCMM is computed by dividing 27 to 44, which is 0.61.

In terms of the metric which measures the cohesiveness of security-critical methods in a given design (i.e., CSCMM), it computes the interactions of methods in a given program with security-critical attributes in the same program taking

		Attributes											
		- Customers. name (0)	+ Customers. dob (2)	+ Customers. nationalID (3)	- Customers. portalPassword (4)	- CreditCards. cardNo (4)	- CreditCards. cardPassword (4)	- DebitCards. cardNo (3)	- DebitCards. cardPassword (4)	- Cheques. chequeNo (1)	- Cheques. chequeAmount (2)	- Loans. interestRate (0)	- Loans. loanAmount (1)
Classes	+ Customers	0	0	3	0	0	0	0	0	0	0	0	0
	+ Credit Cards	0	2	3	0	0	0	0	0	0	0	0	0
	+ Debit Cards	0	2	3	0	0	0	0	0	0	0	0	0
	+ Cheques	0	0	3	0	0	0	0	0	0	0	0	0
	+ Loans	0	0	3	0	0	0	0	0	0	0	0	0
Methods	+ Customer.getName	0	0	0	0	0	0	0	0	0	0	0	0
	- Customer.getDOB	0	2	0	0	0	0	0	0	0	0	0	0
	- Customer.getNationalID	0	0	3	0	0	0	0	0	0	0	0	0
	+ Customer.getPortalPassword	0	0	0	4	0	0	0	0	0	0	0	0
	- CreditCards.getCardNo	0	0	0	0	4	0	0	0	0	0	0	0
	+ CreditCards.getCardPassword	0	0	0	0	0	4	0	0	0	0	0	0
	+ CreditCards.setCardPassword	0	0	0	4	0	4	0	0	0	0	0	0
	+ DebitCards.setCardPassword	0	0	0	0	0	0	0	4	0	0	0	0
	+ DebitCards.verifyCredentials	0	0	0	4	0	0	3	4	0	0	0	0
	+ Cheques.getChequeNo	0	0	0	0	0	0	0	0	1	0	0	0
	+ Cheques.getChequeAmount	0	0	0	0	0	0	0	0	0	2	0	0
	+ Loans.setInterestRate	0	0	0	0	0	0	0	0	0	0	0	0
	+ Loans. getLoanAmount	0	0	0	0	0	0	0	0	0	0	0	1

Figure 4. Banking System Security Matrix

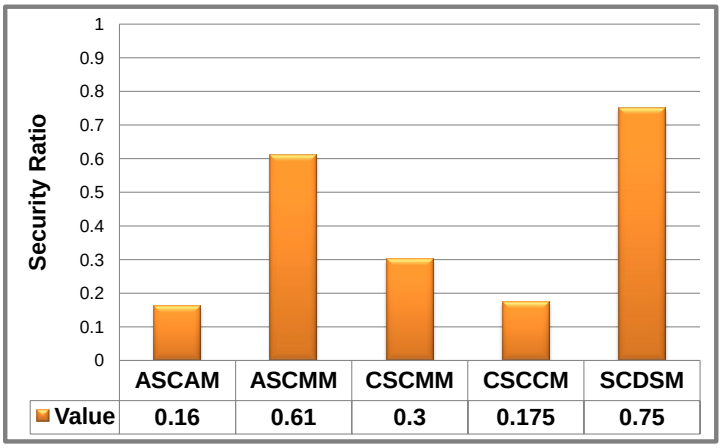


Figure 5. Banking System Results Chart

into consideration the security values of these attributes. The main aim of this metric is to have a low value of cohesiveness as possible in order to meet the requirements of the security design principle of least privilege. From the design used in this context, there are 11 security-critical methods interacting with different security-critical attributes of different security levels. The criticality of interactions is calculated by summing the number of interactions in every class, and then multiplying each interaction by its security value. This result is divided by the number of methods in the same class multiplied by the number of security-critical attributes and by the maximum security value in the model. In the Customers class case, the criticality of interactions of methods in this class with its security-critical attributes is 9. This is extracted from the security matrix in Figure 4 since the criticality of interactions from method `Customers.getDOB` is 2, the criticality of interactions from method `Customers.getNationalID` is 3, and the criticality of interactions from method `Customers.getPortalPassword` is 4. Moreover, the total number of possible interactions from inside the Customers class with its security-critical attributes is 48. The same process has to be done for all classes f the program, then the sum is divided by the total number of classes in that program. Hence, the cohesiveness of the Credit Cards class is 0.33, Debit Cards class is 0.6875, Cheques class is 0.1875, and Loans class is 0.125. Finally, the CSCMM is computed by dividing the sum of the cohesiveness of all classes (i.e., 1.5175) to the number of classes in the design (i.e., 5). which gives 0.3035 and this is the result of the CSCMM for the design shown here.

With regard to the security metric of coupling, it aims

to reduce the number of associations of classes with other security-critical inside a given program. The lower number of such associations, the more secure a program is in terms of the least privilege security design principle. This metric is computed by counting the number of outsider classes which interact with security-critical attributes, and each interaction is multiplied by the security level of these attribute. The security level of each interaction is determined by the security level of the attribute a class is interacting with. Then, this number is divided by the total number of possible associations with security-critical attributes. In Figure 4, there are five security-critical classes which have ten security-critical attributes of different security levels. The number of classes that interact with the security-critical attribute `Customers.dob` is two, and it's security level is two. Hence, the criticality of these interactions when those two numbers are multiplied by each other is four. The number of classes which interact with security-critical attribute `Customers.nationalID` is four, and it's security level is three. Thus, the criticality of these interactions is twelve. The number of classes which interact with security-critical attribute `Customers.portalPassword` is three, and it's security level is four. Thus, the criticality of these interactions is twelve. Finally, the criticality of the total number of links with security-critical classes is 28. The criticality of the possible number of classes that may interact with security-critical classes in this design is 160 (number of classes less by one multiplied by number of security-critical attributes and then multiplied by the maximum security level in this case). Therefore, the CSCCM is computed by dividing 28 over 160, which is 0.175.

The design size metric aims to measure the proportion of security-critical classes in a given program. A lower value of this metric is desirable in order to meet the requirements of the security design principle of reducing the attack surface size. To compute this metric, each security-critical class is multiplied by its security level, and the sum of this is divided over the total number of classes multiplied by the maximum security value in a certain design. The security value of a certain class is the same as the highest security value of the attributes it includes. In the design shown in Figure 4, there are five classes that all of them have security-critical attributes. There are three security-critical classes which their security value is 4 ; Customers, Credit Cards, and Debit Cards. There are one security-critical class which its security value is 2 (i.e., Cheques) and one class which its security value is 1 which is Loans. The sum of these security values is 15. The total security values of all classes in a certain design is computed by multiplying the number of classes (5 in this context) by the maximum security value in a given design (4 in this context). Hence, The SCDSM is computed by dividing 15 over 20 which is 0.75.

D. Results Discussion

Figure 5 shows the results of the five security metrics defined and studied by this paper. This part examines these results in order to show which ones contribute to a more secure design. As shown in the previous section, the results of these metrics vary and this effects the security of the program in reference to the relevant security design principle. For instance, the result of ASCAM is the lowest one among the results of the five metrics defined by this model. Since ASCAM measures the information flow with regard to the accessibility of security-critical attributes in order to meet the requirements of the security design principle of reducing the attack surface size, then this result indicates that it is the best metric which makes the program secure in this regard.

On the other hand, SCDSM measures the information flow of security-critical data in a given object-oriented design with regard to the size of the security-critical classes in that design. Therefore, it is desirable to have as few as possible of security-critical classes to satisfy the security design principle of reducing the attack surface size. However, the result of SCDSM in the case study shown here is the highest one among all of the five metrics, and hence the examined program is the least secure with this regard.

In terms of the metrics which measures the information flow in a given design in order to meet the requirements of the security design principle of granting least privilege, the results of these metrics show that CSCCM is lower than CSCMM. This indicates that CSCCM is more secure than CSCMM in terms of the least privilege security design principle in this case.

VIII. Conclusion

This work has developed a generic model for quantifying security of multilevel object-oriented designs. It concentrates on defining five security design metrics with regard to four different quality design properties for object-oriented programs. These metrics aim to assess the potential information flow of security-critical data within a given program with regard to security design principles of least privilege and attack surface size. This model differs from previous ones as it takes into consideration the criticality of the system's components which is defined by their security levels. Further contribution of this work includes the definition of a security matrix that is developed in order to make it easy for software designers to extract the results of these metrics from the system's UML class diagram. At the end of this paper, a case study of a typical object-oriented program is examined. It has shown in this case study how to apply these metrics, construct the security matrix of that design, and compare the results of these metrics to identify which metrics makes programs more secure than others.

Future extensions of this work include studying the effect of other software quality properties such as polymorphism and inheritance on the security of object-oriented programs. Future work also includes studying these metrics on a lower level of modularity such as at the code-level to identify the impact of these metrics and others on multilevel security-critical programs.

Acknowledgment

I wish to thank the anonymous reviewers for their helpful suggestions. The author gratefully acknowledges Aljouf University for funding this research through grant 35/344.

References

[1] M. Howard and D. LeBlanc, *Writing Secure Code*. Redmond, Wash.: Microsoft Press, 2002.

[2] G. McGraw, *Software Security: Building Security In*. Upper Saddle River, NJ: Addison-Wesley, 2006.

[3] V. B. Livshits and M. S. Lam, "Finding security vulnerabilities in Java applications with static analysis," in *SSYM'05: Proceedings of the 14th conference on USENIX Security Symposium*, (Berkeley, CA, USA), pp. 18–18, USENIX Association, 2005.

[4] S. F. Smith and M. Thober, "Improving usability of information flow security in Java," in *Proceedings of the 2007 workshop on Programming languages and analysis for security*, PLAS '07, (New York, NY, USA), pp. 11–20, ACM, 2007.

[5] G. Smith, *Principles of Secure Information Flow Analysis*, vol. 27, pp. 291–307. Springer, 2007.

[6] Y. Liu and A. Milanova, "Static analysis for inference of explicit information flow," in *Proceedings of the 8th ACM SIGPLAN-SIGSOFT workshop on Program analysis for software tools and engineering*, PASTE '08, (New York, NY, USA), pp. 50–56, ACM, 2008.

[7] I. Chowdhury, B. Chan, and M. Zulkernine, "Security metrics for source code structures," in *Proceedings of the Fourth International Workshop on Software Engineering for Secure Systems*, (Leipzig, Germany), pp. 57–64, ACM, 2008.

[8] J. Bansiya and C. G. Davis, "A hierarchical model for object-oriented design quality assessment," *IEEE Transactions on Software Engineering*, vol. 28, pp. 4–17, 2002.

[9] M. Howard, "Attack surface: Mitigate security risks by minimizing the code you expose to untrusted users," *Microsoft MSDN Magazine*, vol. 11, 2004.

[10] P. K. Manadhata, K. M. C. Tan, R. A. Maxion, and J. M. Wing, "An approach to measuring a system's attack surface," Tech. Rep. CMU-CS-07-146, Carnegie Mellon University, Pittsburgh, PA, August 2007.

[11] P. Manadhata and J. Wing, "An attack surface metric," *IEEE Transactions on Software Engineering*, vol. PP, no. 99, p. 1, 2010.

[12] M. Bishop, *Computer Security: Art and Science*. Boston: Addison-Wesley, 2003.

[13] K. Maruyama, "Secure refactoring - improving the security level of existing code," in *Proceedings of the Second International Conference on Software and Data Technologies (ICSOFT 2007)* (J. Filipe, B. Shishkov, and M. Helfert, eds.), (Barcelona, Spain), pp. 222–229, 2007.

[14] S. Chidamber and C. Kemerer, "A metrics suite for object oriented design," *IEEE Transactions on Software Engineering*, vol. 20, pp. 476–493, 1994.

[15] M. Lorenz and J. Kidd, *Object-Oriented Software Metrics: A Practical Guide*. Englewood Cliffs, NJ: PTR Prentice Hall, 1994.

[16] V. R. Basili, "Gqm approach has evolved to include models," *IEEE Software*, vol. 11, pp. 1–8, 1994.

[17] F. B. e Abreu, M. Goulão, and R. Esteves, "Toward the design quality evaluation of object-oriented software systems," in *Proceedings of the 5th International conference on Software Quality, Austin Texas*, 1995.

[18] B. Henderson-Sellers, *Object-oriented metrics: measures of complexity*. Upper Saddle River, NJ, USA: Prentice-Hall, Inc., 1996.

[19] B. Alshammari, C. J. Fidge, and D. Corney, "Security metrics for object-oriented class designs," in *Proceedings of the Ninth International Conference on Quality Software (QSIC 2009)*, (Jeju, Korea), pp. 11–20, IEEE, 2009.

[20] B. Alshammari, C. J. Fidge, and D. Corney, "Security metrics for object-oriented designs," in *Proceedings of the Twenty-First Australian Software Engineering Conference (ASWEC 2010), Auckland, 6–9 April* (J. Noble and C. J. Fidge, eds.), (California, USA), pp. 55–64, IEEE Computer Society, 2010.

[21] Igor Kotenko and Elena Doynikova, "Comprehensive multilevel security risk assessment of distributed information systems," *International Journal of Computing*, vol. 12, no. 3, pp. 217–225, 2013.

[22] D. Ourston and R. J. Mooney, "Theory refinement combining analytical and empirical methods," *Artificial Intelligence*, vol. 66, no. 2, pp. 273–309, 1994.

[23] J. Jurjens, "Using UMLsec and goal trees for secure systems development," in *Proceedings of the 2002 ACM Symposium on Applied Computing Madrid*, (Madrid, Spain), ACM, 2002.

[24] J. Barnes, *High Integrity Software: The SPARK Approach to Safety and Security*. London, Great Britain: Addison-Wesley, 2003.

[25] P. K. Manadhata, *An Attack Surface Metric*. PhD thesis, Computer Science Department, Carnegie Mellon University, December 2008.

[26] S. Bennett, S. McRobb, and R. Farmer, *Object-Oriented Systems Analysis and Design Using UML*. Maidenhead: McGraw-Hill Higher Education, third ed., 2006.

[27] M. Y. Liu and I. Traore, "Empirical relation between coupling and attackability in software systems: a case study on DOS," in *Proceedings of the 2006 Workshop on Programming Languages and Analysis for Security, Ottawa*, (Ottawa, Ontario, Canada), pp. 57–64, ACM, 2006.

Towards Analytical Modeling for Persuasive Design Choices in Mobile Apps

Hamid Mukhtar
National University of Sciences & Technology (NUST)
44000, Islamabad, Pakistan

***Abstract*—Persuasive technology has emerged as a new field of research in the past decade with its applications in various domains including web-designing, human-computer interaction, healthcare systems, and social networks. Although persuasive technology has its roots in psychology and cognitive sciences, researchers from the computing disciplines are also increasingly interested in it. Unfortunately, the existing theories, models, and frameworks for persuasive system design fall short due to absence of systematic design processes mostly used in the computing domains as well as lack of support for appropriate post-analysis.**

This work provides some insight into such limitations and identifies the importance of analytical modeling for persuasion in mobile applications design. The authors illustrate, using a case study, that appropriate mathematical models can be applied together with user modeling to develop a persuasive system that will allow the designer to consider several design choices simultaneously.

***Keywords*—goal; intent; analytics; modeling; feedback**

I. Introduction

Persuasive technology has its roots in psychology and human behavior and much of the early work done was focused on using websites and mobile phones [1][2][3] as persuasion media; however, as the technology matured and people had increased access to sophisticated devices and systems, it became more natural to design appropriate interactive interfaces and specialized tools as persuasive media. Researchers developed various methodologies [4], models [5] and architectures [6] that could serve as guidelines for designing a wide range of persuasive systems irrespective of the persuasion medium.

Due to such research efforts, persuasive technology has matured in the recent years and it has emerged as a separate field of research. Notably it has found its place in various applications of Human-Computer Interaction (HCI). In fact, persuasive design has recently got much attention and significant role in designing products varying from hand-held devices and household items to software products including websites and mobile applications. However, it was not until very recently when people realized that integrating "analytics" can leverage the persuasion power significantly. That's why, there have been efforts in integrating data logging and analysis in different systems designed with the objective of persuading the users [7][8]. Based upon these trends, the authors present here our view on the need for analytical models in persuasion.

A. The Need for Analytical Models

Fogg's Behavior Model (FBM) [9] defines behavior as a product of three factors: motivation, ability and triggers, and their subcomponents. While in theory this model serves as a good guide for understanding human behavior – and for identifying the target component to work on for inducing the desired behavior – it does not specify a mathematical or computational model or any hint on how to integrate it into such an existing model. This creates a major hindrance in designing persuasive systems that can be evaluated on the basis of expert or machine analysis. Most of the persuasive systems in the literature report about achieving significant change in the behavior of subjects; however, they do not specify to what extent a given component of the FBM was successful on a particular subject. In fact, such analysis cannot be made possible without appropriate analytical models. What if we can design an analytical model of a persuasive system that can not only provide qualitative but also quantitative analysis of the persuasion goal?

At the core of the analytical model, we can have a mathematical or computational model appropriate for each application scenario where persuasion is needed. Having a separate mathematical model for each persuasive situation provides the ability to fine tune the various characteristics associated with persuasion. It also implies that the designer has the ability to adopt the persuasion strategies to individuals or classes of individuals rather than applying the same strategies indiscriminately to all users.

This paper approaches the domain of persuasive technology from data analytics point-of-view with two main objectives. First, the authors want to emphasize that by integrating analytical models as part of the persuasive design leverages the persuasion process and will lead to enhanced outcomes. Second, the authors want to present an analytical model and, subsequently, describe how we can represent a certain human behavior as a mathematical model for persuasion. The author will also describe how this basic model allows persuasion to be personalized to each person's target behavior based upon their past performance.

This article presents the preliminary work on the importance of analytics in persuasion and is based upon the learning from the previous work on the design and development of persuasive healthcare applications [10][11][12]. Nevertheless, it also draws on the experiences of the work done by others in the domain of persuasive technology and pointers to them are provided throughout the article.

The rest of this article is organized as following. Section II provides the background and overview of the related work. Section III identifies some persuasive design elements which

form the basis for the definition of the analytical model presented in Section IV. Section V describes a case study developed using the analytical model. Section VI concludes this article with some future directions.

II. Background and Related Work

The various studies related to Behavior Change Support Systems (BCSS) [13][14] and particularly the Persuasive System Design (PSD)[5] process provide some model for designing persuasive systems. However, while such model may provide a strong theoretical background to understand the broad spectrum of persuasion, they are not sufficiently detailed to help a designer think in more concrete terms. This has resulted in a large gap between the theory and practice of persuasive technology.

For example, most of the work in persuasive technology start with a focus on Fogg's Behavior Model [9] and designing persuasive strategies for increasing one or more of its ingredients, namely, motivation, ability, or trigger. However, this is where the theory disappears and the focus is turned towards design elements of persuasion related to user interaction and perception. This is evident from the numerous research efforts such as [2][3][7][15]. In contrast, the author believes that by considering some analytical model, it becomes more instructive to consider various additional factors that are outlined below.

Fogg's behavior grid [16] identifies 35 ways a behavior can be changed or induced. The behavior grid categorizes the behavior on the basis of two dimensions: type of behavior change and time or schedule. Accordingly, a target behavior may fall into one of these 35 categories represented by a cell in the grid. Although the behavior grid does not restrict that a given behavior may fall into one category only, it also does not give any clue on mapping a target behavior on more than one cells over an extended period of time. Due to the absence of a formal mechanism, such mappings cannot be done systematically.

Fogg and Hreha [17] simplified the behavior grid to 15 states. The behavior axis has three types: dot (one-time), span (for a set duration) and path(permanent change). Each of these behaviors has five flavors: green, blue, purple, grey and black characterizing if a behavior is familiar or not and whether the intention is to start, increase, decrease or stop the behavior. Of these 15 behavior targets, the designer will spot one and persuade the user for that particular behavior duration and flavor. This will be in one of the 15 cells of the grid where the desired behavior duration and flavor intersect.

Although the current research does not concern the behavior grid, the interest here is in defining models that may place a user into one of the cell of the grid based on certain criteria. Then the user can be moved from cell to cell as his/her behavior is refined. During this process, certain elements from the area of gamification are borrowed, such as the notion of levels or progress. The objective is to justify the fact that by integrating certain level of analytics one can enhance the persuasion process and greatly help the designer identify a number of relevant persuasion strategies. Also, by integrating the concepts of context and user preferences from the lessons learned in the domain of ubiquitous computing, persuasion can be made more meaningful to the user.

III. Designing for Analytics in Persuasion

Recently, business intelligence and machine learning have gained much attention due to increased importance and capabilities of analytics in many fields. Thus, recently many persuasive technologies have been designed on top of analytics [2][8][18]. But in the absence of some analytical model, such efforts are not reusable in different contexts or for different problems. Based upon our previous experiences of developing persuasive systems of different nature, some common design factors have been identified and combined into a unified analytical model for persuasion.

In the remaining part of this section, these factors are identified and then in the next section the model is described. The following scenario is adapted for the purpose of explanation.

1) Example Scenario: Consider the problem of environmental hazards or climate effects resulting from excessive usage of electricity which is mostly generated from natural sources such as coal, petroleum or nuclear energy; all of them are exhaustive in nature. To minimize the environmental impacts due to electricity usage, a non-governmental body decides to persuade people to reduce electricity consumption and to move towards renewable source of energy such as solar energy. Thus, the design will include to deal with two behaviors either simultaneously or independently one of them. Keeping the persuasion goal in mind, we now describe how a persuasion designer can design some viable solution for inducing the desired target behavior in users.

A. Difference between Persuader and User Goals

Clearly, the intent of the persuader is to protect the environment and one possible way is by reducing the electricity consumption. Having the clearly defined goal of persuader, the designer may design persuasive strategies related to the Grey Path cell in the Fogg's behavior grid: *decreasing a behavior that is always performed*. For example, the designer may be tempted to start by persuading the consumers that reducing electricity consumption may be a noble cause of saving the humanity. A persuasion strategy would be informing the users about these negative effects so that they can make informed decisions about reducing electricity consumption.

Although this is a direct translation of the persuader's intent into a persuasion objective, it cannot be said for sure if it can be mapped to most of the users' intent. In other words, the persuasion strategy *reduce usage to save environment* may represent the persuader's intent but not necessary the user's intent. Also, the desired behavior may not be just about reducing a particular habit but alternative behaviors should also be considered.

For a persuasion strategy to be successful, the designer has to come up with a strategy or set of strategies that match user's goal. The challenge here lies in finding the user intent and then doing a translation of the persuader's intent into the user's intent. For example, the stated persuasion strategy of saving environment can be redefined as *reduce electricity usage to save money*. This strategy can be then put into practice by

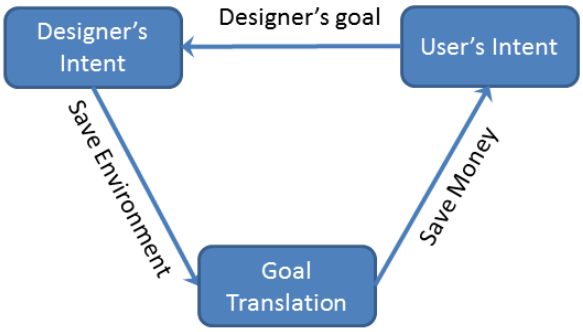


Fig. 1. Translating from Designer Goal to User Goal

raising awareness about the relationship between number of electric units consumed and the resulting saving in money. Here one must consider the fact that this will change the design of the application significantly and introduce some sort of model to take care of money saving but it will also result in a thought process of doing the translation. Such translation may be done easily for both mathematical and persuasion point of view or it may require detailed knowledge of the domain (e.g., environmental factors) by the designer. This concept is explained in 2.

However, not all users may have the same intent and this strategy may not be helpful in changing their motivation or attitude for electricity usage. The reason could be cheap electricity tariffs or the lack of interest by the user in saving money. To solve such differences, one needs to consider different users' different intents, and for that persuasion facets are used.

B. Persuasion Facets

When building a persuasive system, the designer is mostly focused on achieving *one* particular goal (of achieving the specific, desired behavior). For effective persuasion, however, multiple facets of the persuasion process are to be considered. For example, for reducing electricity consumption one persuasion facet is to emphasize on saving money. Another facet can be emphasizing usage of alternative, renewable energy means such as solar-energy-based electricity generation and usage. Yet another facet may be to persuade the users to use smart meters that can identify to the users certain appliances whose usage can be tracked and adjusted according to users' need, e.g., heating systems whose thermostats can be optimized for better economy of electricity.

The designer of a persuasive system should identify different possible facets and then apply persuasion according to the selected facet. Each of the persuasion facets has to be mapped to some user intent clearly. Nevertheless, this mapping cannot be done without considering some additional factors: persuasion context and user preferences.

1) Persuasion Context: Given that persuasion has various facets, persuasion context plays an important role to assist the designer in choosing the most appropriate facet according to a given user's intent. In other words, given different facets and user intents, the persuasion context serves as a mapping function from some facet onto some intent. In literature, the term context is defined as any information that can be used to characterize the situation of an entity [19]. Below are given consider some examples of context.

Having understood the various facets and user intents in the persuasion problem of reducing electricity consumption, a designer can identify a number of entities as context. For example, considering current season or weather as context will allow understanding the appropriate user intent that will in turn determine the appropriate persuasion facet. When the weather or season in the context is cold, it will not be appropriate to consider the option of reducing heater consumption as an economy measure. On the other hand, on warm and sunny areas, it is more appropriate to suggest using alternative means such as solar energy. Other persuasion contexts for this example may include location of usage (e.g., home, office, hotel, or other public place) and time (morning, evening, or night). More sophisticated context entities such as user's education, income, family size, age group, habits or routine, etc. may also needed to be considered in advanced cases.

2) User Preferences: Another factor or function that will help a designer in mapping a persuasion facet onto a user intent is to consider user preferences. To understand the mapping function of user preferences, let's assume that some users may never like the idea of investing into or thinking about alternative energy means including solar energy, because of little motivation or ability. Such likings or dislikes of the user must be considered by the designer carefully because user preferences may be a stronger function than the context function. In case of a conflict between the two or even if a tie occurs between them the user preference will override the context function. Thus, even in warm sunny areas, such users may not be persuaded for using alternative energy means.

C. Feedback is important in Persuasive Design

How does a designer know which of the facets the user is interested in? One possible way is to ask the user in the beginning. However, this approach is not flexible and also some users may not be sure about their goals. The designer may need to determine this indirectly using user's behavior. For this purpose, a feedback mechanism is involved from the system or application to the user and vice-versa. So initially, the system has no knowledge about the user but as the system learns about the user as part of the user's interaction with the system, the latter provides more and more feedback in that particular aspect. For example, having shown the user all the facets towards meeting the goal, the system may monitor user's interaction with the provided feedback (electricity consumption, meter's reading, etc.) and will adapt the interface showing more detailed related to the facet in which the user is interested bringing more of persuasion in it.

D. Persuasion has Several Levels

The author proposes that to achieve a target behavior, a user may sometimes need to go through several levels (or cells in the behavior grid) before arriving at the final level of the target behavior (unless the behavior is so easy to do or obvious for the user that he take a direct start in the final level). This is where the notion of levels is defined in persuasion similar to levels in computer or video games.

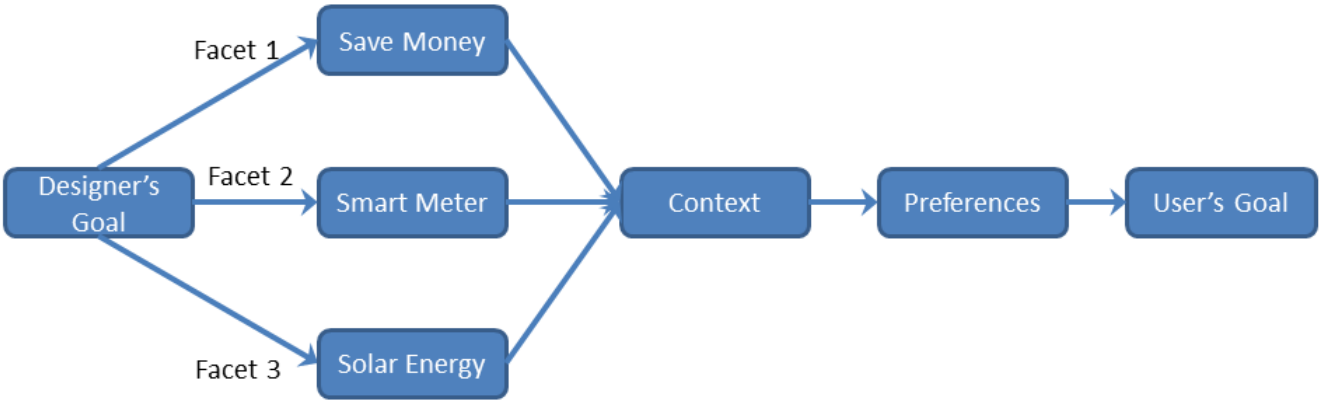


Fig. 2. Persuasion Facets as Alternative Translations into User Goals

Depending upon the persuader's goal, user's goal, persuasion facets and translation functions, it is the job of the persuasion designer to identify the levels required in achieving the target behavior. Although in the current work does not consider any mechanisms on how a designer could do this task, it can be said safely that one possible approach will be to consider the Fogg's notion of tiny habits[1]. Stated simply, the designer may start with very easy to do behaviors that can be converted into habits with least of efforts. As the user establishes himself in a given level (or cell), he may be moved on to the next level with relatively difficult target behavior to do. This may continue from level to level until the desired target behavior is achieved in the final level.

For the given example, the designer may initially persuade the user to install solar energy panel for one room or device: a one-time behavior. Once this is achieved, the user may be persuaded to install the panels for the whole house or additional rooms. This will be successful only if the user sees the benefit of the single installation which can be done by providing feedback to the user on comparative saving of the two energy-consuming mechanisms.

E. Persuasion is a Process

Instead of devising one strategy, the designer may need to identify a number of levels that will be carried out by user as a "process" before doing the desired target behavior. This may apply to even the act of performing a new behavior one-time only. For instance, in the current example of reducing electricity consumption, the most useful behavior would be to use alternate energy means. This designer goal can be achieved efficiently if the user performs a one-time behavior of investing in the solar energy panels. However, to do that one time behavior, which is quite difficult for the user in terms of motivation, the user may first need to go through other persuasion levels that will increase the motivation of the user sufficient enough that a simple trigger may lead him to perform the behavior of investing in alternative means. This process has been depicted in the model described next.

[1]http://tinyhabits.com

IV. Towards Analytical Model for Persuasion

In the previous section, various issues were discussed that need to be addressed in order to introduce analytics in the persuasive design process. In this section, building on the previous concepts, it is explained as how they can be useful in the development of appropriate analytical and mathematical models.

A. The Core Model

Figure 3 shows a model for an analytical persuasive system after considering these factors. The important thing to note is that this model is cyclic by nature. It has been modeled as a continuous process consisting of some iterations. During each iteration the user's behavior is observed or monitored through his actions or activities while considering the context and user preferences. At the end of the iteration, an assessment is made about the user's actual behavior. While monitoring for the behavior, an analysis is performed in tandem and based upon changes in the context, the most appropriate persuasion facet is applied.

Note that the model contains two cycles. The micro-cycle is used to choose an appropriate facet while the macro-cycle allows the persuader to intervene and make an informed assessment to change the persuasion level. However, this intervention may not be required or desirable in most cases and many systems may do well without this.

Such models must define some way of tracking or measuring the user behavior so that it can be identified in which persuasion level to place the user at a certain time. In addition, they should facilitate us in choosing the appropriate persuasion facet by provide insight into the context and user preferences functions. This is where the importance of user profiling comes into play.

B. User Profiling

Data plays an important role in analytics and no analytics are possible without having relevant data. That is why the core model contains a profiling component which serves as initial data gathering mechanism for identifying user preferences. However, the process of data gathering continues in each

Fig. 3. A model for analytical persuasion

iteration in the form of behavior monitoring. Without data gathering, one will not be able to infer context and choose the appropriate persuasion facet. Such data logging techniques are at the heart of various ubiquitous systems [2][3] and have been used for persuasion [7] as well. It is through the process of user profiling that the designer can infer about user's goal to help them in choosing the appropriate persuasion facet.

C. User Modeling

To selectively log data that can be used for persuasion, the user's actions must be understood. This requires the development of appropriate user models specific to each problem. Developing a user model is challenging task as it requires considering some detailed analysis of user's explicit or implicit actions [20][21]. This user model will be used for defining the profiling stage of the core model as well as for specifying what to monitor in each iteration. Most of the times, such models will be a computer representation of the user alongside some mathematical equations to allow reasoning on the data [22][23].

V. Case Study: Sedentaware

In this section, the author explains how analytical modeling can be applied for persuasion in some previous work [10]. Sedentaware is mobile application to raise awareness about sedentary behavior in users and to persuade them to do physical activities, if and only when needed. While designing the application, we had to consider several analytical issues as identified above and briefly described here.

Sedentaware uses appropriate alarms (or triggers) to motivate users to exercise. However, unlike many of the similar approaches developed previously, the app uses a mathematical model to determine the appropriate moment for alarm. This concept is depicted in figure 4, adapted from our previous work [10] for four different users with varying levels of behavior. As it can be observed, user1 is active 50% of the time and gets no alarm (or trigger), because he does not need one. As the sedentary behavior of the users increase, they get more and more alarms. In other words, based on the mathematical model the persuasion is adaptive to individual user.

Moreover, a number of persuasion facets were identified. Figure 5 shows four different strategies with the intention of motivating the users to participate more in active behavior. First, in figure 5(a) provides a glimpse of user's current progress. A user is motivated if he sees that he is lacking in progress at a given moment. Second, figure 5(b) provides a daily statistics view as a graph of percentages allowing the user to reflect on his progress over some time. Third, figure 5(c) provides a weekly progress view as a different graph as minutes of different activities performed. This graph provides another facet or window to view their behavior differently. Finally, a fourth persuasion facet is provided in the form of instinct cues to the user. The figure shows an icon on the top left of the mobile screen whose color provides the user a cue on how well he is doing. In this example, these facets are just looking at the same view from different windows. However, the same mathematical model can be used to introduce ranking of users in his social network resulting in a completely different facet of the persuasion strategy of 'competition' with the peers.

Behind these different persuasion facets exist the various elements of a mathematical model that computes the ratio of user's active behavior (walking, running, etc.) to the sedentary behavior. The behavior is detected using activity recognition on user's mobile phone. Interested readers are referred to [10] for further details on development of the model, the data logging procedure used in the application as well as the evaluation by test users. Note that the application translates the persuader goal of prevention of sedentary behavior into the activeness goal of the user. The context of the user (location and time) is also considered. Finally, as the user is adopting the desired behavior, he/she is also progressing in the levels as shown in figure 6.

As shown in figures 4-6, this model is not only used in designing various persuasion facets but also for evaluating the user's performance by them or even by the designer, assuming the designer has access to the data using appropriate means.

VI. Conclusions

In this article, we described the use of analytics in the field of persuasive technology for better design choices to the persuasive system designers. We described several factors to consider when designing for analytics in persuasion and proposed a model that considered persuasion as a process. Using a mobile application, called Sedentaware, as an example for preventing sedentary behavior by motivating users to carry out physical activities, we explained how our model can be used alongside a mathematical model to create enhanced persuasive technologies. The current work identified an important limitation of persuasive technology theory that it lacks the detail required by the designers to carry out systematic process of designing persuasive technologies.

This work will be extended to include further concepts from the persuasion and behavior theory, particularly the work related to Behavior Change Support Systems (BCSS). We also need to identify how different persuasion intents can be mapped to Fogg's behavior grid. It will only be at that time when we will be confident about enhanced role of researchers from the field of computer sciences and relevant domains in persuasive technologies.

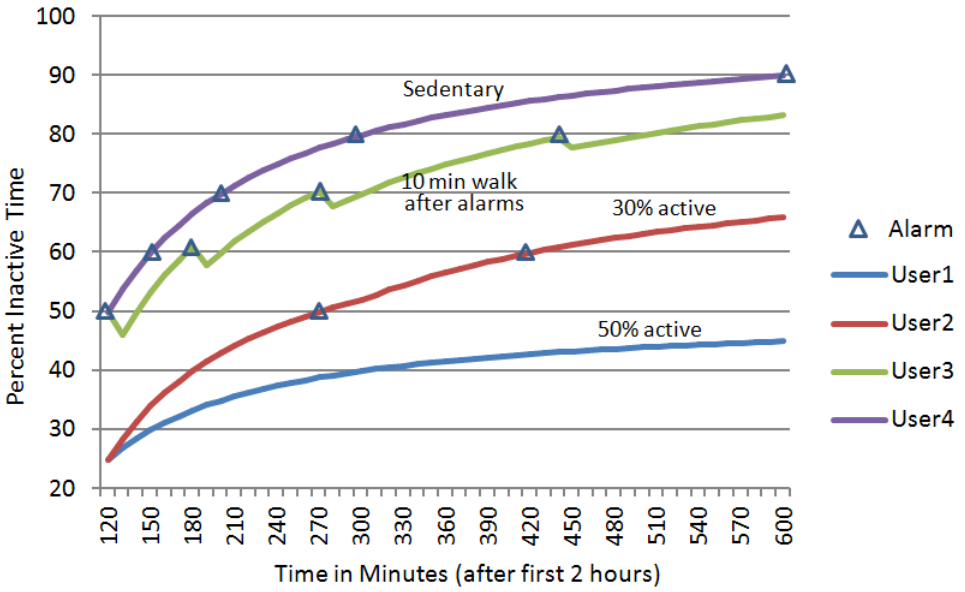


Fig. 4. Results of simulation showing sending of adaptive reminders to various users with different sedentary behavior. Active users receive fewer and less frequent reminders as compared to sedentary users.

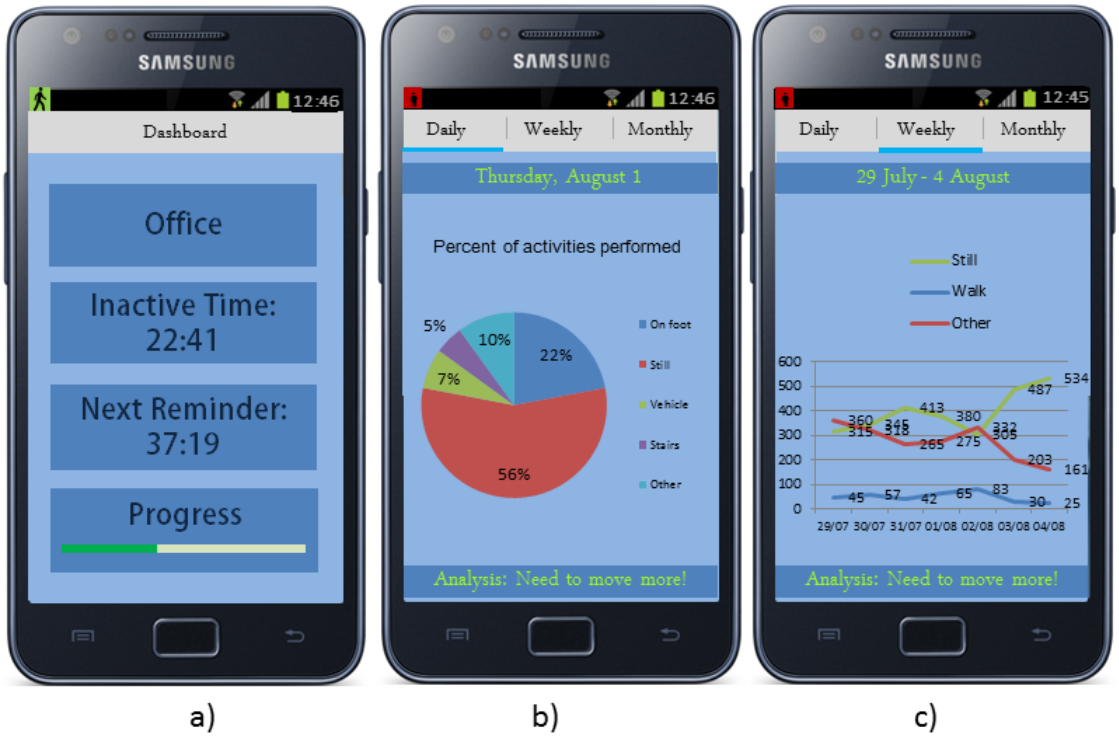


Fig. 5. a) Activity Dashboard b) Summary of Daily Activities c) Weekly Summary of Activities. Both the notification bar and the application's screen reflect the change using various colors

Fig. 6. Feedback and Levels

REFERENCES

[1] B. Fogg, "Persuasive technology: using computers to change what we think and do," *Ubiquity*, vol. 2002, no. December, p. 5, 2002.

[2] H. Kimura, J. Ebisui, Y. Funabashi, A. Yoshii, and T. Nakajima, "idetective: a persuasive application to motivate healthier behavior using smart phone," in *Proceedings of the 2011 ACM Symposium on Applied Computing*, ser. SAC '11, 2011, pp. 399–404.

[3] S. van Dantzig, G. Geleijnse, and A. T. van Halteren, "Toward a persuasive mobile application to reduce sedentary behavior," *Personal and Ubiquitous Computing*, pp. 1–10, 2011.

[4] B. Fogg, "Creating persuasive technologies : An eight-step design process," in *Persuasive '09*, 2009, p. 16.

[5] K. Torning and H. Oinas-Kukkonen, "Persuasive system design: state of the art and future directions," in *Proceedings of the 4th International Conference on Persuasive Technology*. ACM, 2009, p. 30.

[6] T. Alahäivälä, H. Oinas-Kukkonen, and T. Jokelainen, "Software architecture design for health bcss: case onnikka," in *Persuasive Technology*. Springer, 2013, pp. 3–14.

[7] S. M. Kelders and J. E. L. van Gemert-Pijnen, "Using log-data as a starting point to make ehealth more persuasive," in *Persuasive Technology*. Springer, 2013, pp. 99–109.

[8] D. Pavel, V. Callaghan, and A. K. Dey, "Looking back in wonder: How self-monitoring technologies can help us better understand ourselves," in *Intelligent Environments (IE), 2010 Sixth International Conference on*. IEEE, 2010, pp. 289–294.

[9] B. Fogg, "A behavior model for persuasive design," in *Proceedings of the 4th international Conference on Persuasive Technology*. ACM, 2009, p. 40.

[10] H. Mukhtar and D. Belaïd, "Using adaptive feedback for promoting awareness about physical activeness in adults," in *Accepted for publication in 10th IEEE International Conference on Ubiquitous and Intelligent Computing (UIC) 2013*, 2013.

[11] H. Mukhtar, A. Ali, S. Lee, and D. Belaïd, "Personalized healthcare self-management using social persuasion," *Impact Analysis of Solutions for Chronic Disease Prevention and Management*, pp. 66–73, 2012.

[12] H. Mukhtar, A. Ali, D. Belaïd, and S. Lee, "Persuasive healthcare self-management in intelligent environments," in *Intelligent Environments (IE), 2012 8th International Conference on*. IEEE, 2012, pp. 190–197.

[13] H. Oinas-Kukkonen, "A foundation for the study of behavior change support systems," *Personal and Ubiquitous Computing*, pp. 1–13, 2012.

[14] ——, "Behavior change support systems: A research model and agenda," in *Persuasive Technology*. Springer, 2010, pp. 4–14.

[15] S. Consolvo, J. Landay, and D. McDonald, "Designing for behavior change in everyday life," *IEEE Computer*, vol. 405, pp. 100–103, 2009.

[16] B. Fogg, "The behavior grid: 35 ways behavior can change," in *Proceedings of the 4th international Conference on Persuasive Technology*. ACM, 2009, p. 42.

[17] B. Fogg and J. Hreha, "Behavior wizard: a method for matching target behaviors with solutions," *Persuasive Technology*, pp. 117–131, 2010.

[18] I. Li, A. K. Dey, and J. Forlizzi, "Using context to reveal factors that affect physical activity," *ACM Transactions on Computer-Human Interaction (TOCHI)*, vol. 19, no. 1, p. 7, 2012.

[19] A. K. Dey, "Understanding and using context," *Personal and ubiquitous computing*, vol. 5, no. 1, pp. 4–7, 2001.

[20] S. McBurney, N. Taylor, H. Williams, and E. Papadopoulou, "Giving the user explicit control over implicit personalisation," in *Procs. of Workshop on Intelligent Pervasive Environments (under AISB09), Edinburgh, Scotland*, 2009.

[21] Z. Jrad, M.-A. Aufaure, and M. Hadjouni, "A contextual user model for web personalization," in *Web Information Systems Engineering–WISE 2007 Workshops*. Springer, 2007, pp. 350–361.

[22] J. Fink and A. Kobsa, "User modeling for personalized city tours," *Artificial intelligence review*, vol. 18, no. 1, pp. 33–74, 2002.

[23] D. Heckmann, *Ubiquitous user modeling*. IOS Press, 2005, vol. 297.

Computer Science Approach To Philosophy: Schematizing Whitehead's Processes

Sabah Al-Fedaghi
Computer Engineering Department
Kuwait University
Kuwait

***Abstract*—Diagrams are used in many areas of study to depict knowledge and to assist in understanding of problems. This paper aims to utilize schematic representation to facilitate understanding of certain philosophical works; specifically, it is an attempt, albeit tentative, to schematize A. N. Whitehead's ontological approach. It targets professionals and students in fields outside of philosophy such as computer science and engineering, who often look to sources in philosophy for design ideas or for a critical framework for practice. Yet students in such fields struggle to navigate thinkers' writings. The paper employs *schematization* as an apparatus of specification for *clarifying* philosophical language by describing philosophical ideas in a form familiar to computer science. The resultant high-level representation seems to be a viable tool for enhancing the relationship between philosophy and computer science, especially in computer science education.**

Keywords*—A. N. Whitehead; schematization; metaphysical ontology; diagrammatic representation; flow

I. Introduction

This paper aims to employ diagrammatic representation to facilitate understanding of philosophical works; specifically, it is an attempt, albeit a tentative one, to schematize A. N. Whitehead's ontological approach. It targets professionals and students in fields outside of philosophy such as computer science and engineering, who often look to sources in philosophy for design ideas or for a critical framework for practice. According to Schwill [1], "It is necessary that students obtain a sketch of the fundamental ideas, principles, methods and ways of thinking of computer science. Only these fundamentals seem to remain valid in the long term and enable students to acquire new concepts successfully during their professional career."

Yet students in such fields struggle to navigate thinkers' writings. The philosophy of Alfred North Whitehead is described as "arguably among the least understood and appreciated works of the Twentieth Century" [2]. For example, it is repeatedly stated that Whitehead's *Process of Reality* [3] is "a complex and a difficult book" [4]. Stengers [5] describes the book as "a text which has repelled so many readers" by its "astonishing difficulty."

Nevertheless, Whitehead put wide-ranging knowledge of science, history, philosophy, mathematics, and mathematical physics all together "in a way which seemed to many people to make the twentieth century world intelligible, that attracted readers far beyond the usual audience for philosophy" [6]

The importance of … Whitehead does not lie in simply learning his philosophy, adopting his terminology, and applying it to a set of research problems. Instead, the demand is to rethink the conceptual and practical procedures and problems that we have inherited and dwell within. [7]

The focus on Whitehead's work "is warranted by both his impacts on science and the current relevance of his work for inspiring new approaches to numerous topics in science and the humanities" [8].

Accordingly, in addition to benefiting students in computer science and engineering, this paper utilizes schematization to achieve certain aims, as follows:

A. Schematization method

The paper employs *schematization* as an apparatus for specification instead of, say, a written description. Schematization is utilized for the purpose of *clarifying* philosophical language; i.e., specifying such language in a form familiar to computer science.

Schematization is conceptualized in this paper:

- as an abstraction
- as a mechanism, machine, process
- as a diagrammatic representation
- as a representational map (e.g., city traffic map)
- as a representation of (sequence) action scene
- as a construction tool for a conceptual model
- as an engineering-like schema with generalization

Accordingly, representing the user as a stickman, as in UML *use diagrams*, is not the type of schematization applied in this paper. As we will see, the *user* in the proposed model is depicted in terms of a sphere that includes five stages such as creating (e.g., an order), and receiving (e.g., an invoice).

Many scientific fields use diagrams to depict knowledge and to assist in understanding problems. "Today, images are … considered not merely a means to illustrate and popularize knowledge but rather a genuine component of the discovery, analysis and justification of scientific knowledge" [9]. "It is a quite recent movement among philosophers, logicians, cognitive scientists and computer scientists to focus on different types of representation systems, and much research

has been focused on diagrammatic representation systems in particular" [10].

In philosophy, images and diagrams are old subjects. Plato's allegory of the cave depicts knowledge configurations. "The diagram functions as an instrument of making evident the structure of ontology and epistemology… [Descartes made] two-dimensional geometric figures and linear algebraic equations mutually transferable" [9].

B. Aims of the paper

This paper explores the diagrammatic representation to be applied in schematizing *flows* and structured *events* in Whitehead's ontology. Advantages of the resultant diagrams include a more concrete description, from the viewpoint of computer science, of philosophical concepts and problems, and new variations in consideration of these concepts and how to reflect about them.

However, a more ambitious aim of the paper is to explore the possibility of incorporating Whitehead's philosophy into computer science. *Traditional* philosophical discussion seldom receives more than academic interest in computer science; nevertheless, several interesting philosophical problems have attracted attention, including the ontological position of software, intellectual property rights, privacy issues, and problems in artificial intelligence.

Many computer science theories are related to philosophy, e.g., object-orientated programming and concurrency. Philosophy can use computer science as a vehicle for "possible 'experimental Philosophy' which is able to provide practical tests for different philosophical ideas" [11]. The so-called philosophy of computer science is said to be concerned with conceptual issues that arise from reflection on the *nature* of computer science [12].

Nevertheless, one of the underlying motivations in this paper is the need in computer science for a broader connection to philosophy. This paper asks two main questions about philosophy in relation to computer science:

- How to *use* philosophy in computer science?
- How to *read* philosophy in computer science?

The following case exemplifies the first question.

Ventura [13] asks, How do we perceive an object identity along time in the context of software applications? And, how to relate object-oriented software to philosophical theories?

A specific Client object in … [an] application will contain the most updated contact information, and, therefore, it won't contain, for instance, the previous phone number of the contact person. It doesn't mean, of course, … But still, the "historical" object itself … will not contain the history of the selling activity that it documents. ... When we talk about an object's history, we actually deal with one of the most profound issues of metaphysics … the question of object continuity. [12]

Ventura [13] traces efforts to resolve the problem to two rather different philosophical approaches: the Perdurantism Approach and the Endurantism Approach. This example illustrates the meaning of *exploring the possibility of incorporating Whitehead's philosophy into computer science*.

Facilitating a broader movement in this direction leads to consideration of the second question mentioned above, how to read philosophy in computer science? *Schematization* is one of the main tools used in computer science to "read" a system, e.g., flowcharts, UML, and SysML; accordingly, the focus in this paper is on schematizing Whitehead's processes.

Given the number of papers produced in the past fifty years on Whitehead's writings, a separate review of literature is not necessary. Instead, quotations about a notion under discussion will be interwoven into the appropriate text in the paper. We assume that the reader is at least slightly familiar with Whitehead's concepts and knows basic philosophical terminology such as the term ontology. Additionally, because of space limitation, only some of Whitehead's ideas are discussed to demonstrate the viability of the diagramming method.

In the next section, the paper begins by reviewing the modeling tool, called the Flowthing Model (FM) [14-16], to be used in interpreting Whitehead's notions through diagrammatic representations. Generally speaking, the word *model*, as used here, is a computer science term that refers to an abstract representation developed as a means of communication among *stakeholders* when building a complex system. FM has been adapted for several applications, and the *Earth seasons* example given here is a another new contribution.

II. Flowthing Model

The Flowthing Model (FM) was inspired by the many types of *flows* that are found in diverse fields, including information flows, signal flows, and data flows in communication models. This model is a diagrammatic schema that uses *flowthings* to represent a range of items, for example, electrical, mechanical, chemical and thermal signals, blood, food, concepts, pieces of data, and so on. Yet, *flow* in FM does not designate only mobility; rather, it encompasses creation and transformation.

Flowthings are defined as *what can* be created, released, transferred, processed, and/or received (see Fig. 1). In the field of ontology, a flowthing can be called an object (substance ontology) or an actual entity (process ontology). Hereafter in the paper, flowthings are referred to as *things*. The (abstract) machine shown in Fig. 1 is a generalization of the typical input-process-output model used in many scientific fields (Fig. 2).

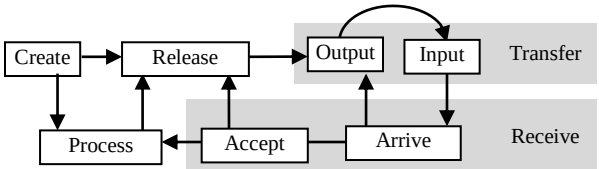


Fig. 1. Flow machine

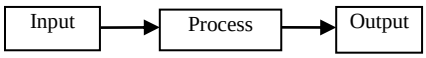


Fig. 2. Input-process-output model

FM depicts *flow* by using *flow machines* (Fig. 1) comprising up to six stages (states). The term *machine* is used here in the sense of *system* or *organism*. The machine is the conceptual fiber used to handle flowthings (to *change* them through stages) from inception or arrival to de-creation or transmission to outside the system. Hereafter, flow machines will be referred to as *machines*. Machines form the organizational structure of whatever is described. These machines can be embedded in a network of assemblies and hierarchies called *spheres*.

The stages in Fig. 1 can be described as follows:

Arrive: A thing reaches a new machine (curved arrow in Fig. 1).

Accepted: A thing is permitted to enter the machine. If arriving things are always accepted, *Arrive* and *Accept* can be combined as a *Received* stage.

Processed (changed): The thing goes through some kind of transformation that changes it without creating a *new* thing. The change may trigger the creation of new flowthings.

Released: A thing is marked as ready to be transferred outside the machine.

Transferred: The thing is transported somewhere from/to outside the machine.

Created: A new thing is born (created) in a machine and its sphere. It is the *becoming* of that which has no prior being (appearance of a new thing in the sphere), e.g., a new actor appears in a scene, not as a person coming from outside, but by suddenly being in the spotlight on a previously dark place on the stage.

In general, a flow machine is thought to be an abstract machine that receives, processes, creates, releases, and/or transfers things. The stages in this machine are mutually exclusive for atomic flowthings; that is, they are indivisible, nor do they spread over two stages. Suppose that a *car* is being created (manufactured); it cannot be released from the assembly line before the end (e.g., say at the stage where it is just a body with some electrical wiring). It must become *a car* and fulfill certain conditions before it can be released.

An additional stage of *Storage* can also be added to any machine to represent the storage of things (memory); however, storage is not an exclusive stage because there can be *stored processed* things, *stored created* things, etc.

FM also uses the notions of *spheres and subspheres*. These are the network environments and relationships of machines and submachines. Multiple machines can exist in a sphere if needed. A sphere can be a person, an organ, an entity (e.g., a company, a customer), a location (a laboratory, a waiting room), a communication medium (a channel, a wire). A flow machine is a subsphere that embodies the flow; it itself has no subspheres.

FM also utilizes the notion of *triggering*. Triggering is the activation of a flow, denoted in FM diagrams by a *dashed arrow*. It is a (causative) dependency among flows and parts of flows. A flow is said to be triggered if it is activated by another flow (e.g., a flow of electricity triggers a flow of heat), or activated by another point in the flow. Triggering can also be used to initiate events such as starting up a machine (e.g., by remote signal). Multiple machines captured by FM can interact by triggering events related to other machines in those machines' spheres and stages.

Example of FM representation: According to Whitehead [17], the *permanence of things* is exemplified by physical things such as the solid Earth, mountains, stones, and the Egyptian Pyramids; however, the Earth *flows* around the sun, as depicted in Fig. 3. Fig. 4 shows the corresponding FM representation. The FM representation indicates that each season is actually a sequence of transferring, receiving, processing and releasing of the flowthing Earth.

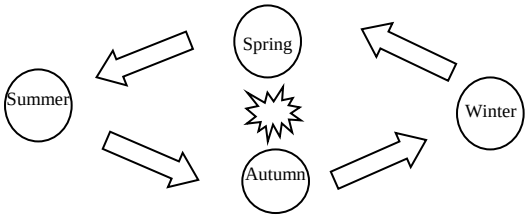


Fig. 3. Earth flows around the sun (redrawn from [18])

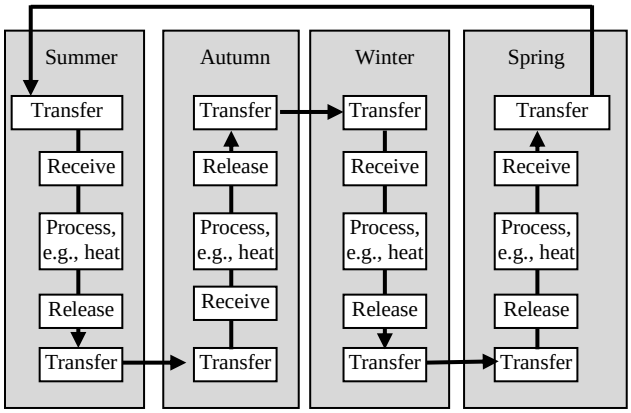


Fig. 4. Earth as a flowthing

Note that in Fig. 4, each season should have been modeled as a sphere that includes the Earth machine, as shown in Fig. 5; however, for simplicity's sake the machine and sphere boxes are represented by one box in Fig. 4.

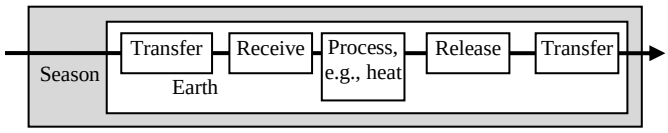


Fig. 5. Earth machine (system/organism) in the Season sphere

III. WHITEHEAD'S ONTOLOGY

Whitehead [17] refers to Heraclitus's statement that *all things flow* as the first generalization "around which we must weave our philosophical system." According to Whitehead [17], a rival antithetical notion (*substance* ontology) can be given for *all things flow* by pointing out the *permanence of things*. According to the ontology of material substance (from Democritus to Newton), everything can be reduced to basic

elements that interact mechanically and lack interiority themselves.

Substances are material things… comprising independent parts, each adapted for a specific function and moving in a specific manner... In substance ontology, processes rearrange matter and, since matter lacks a subjective nature, processes happen *to* matter. [19]

On the other hand, *process* ontology considers process a fundamental descriptor of reality [19]. A *process* indicates a mode of change: "Coordinated group of changes in the complexion of reality, an organized family of occurrences that are systematically linked to one another either causally or functionally. It is emphatically not necessarily a change in or of an individual thing, but can simply be related to some aspect of the general 'conditions of things'. ... Processes are existentially fundamental; substance is mere appearance" [20]. Processes are partly self-determining (subjective), and can enter into relation with other processes [19]. "They are not themselves temporal. Each one is an indivisible epoch having no internal temporal phase" [21].

A. Actual entities

Dynamic reality exists in terms of *actual entities*. Actual occasions (events) are the basic units of process or *becomingness* [22]. *Becoming* refers to the process of emerging as a thing. Here, *actual* contrasts with *potential*. "To be actual is to be a process" [23].

The world is certainly an ongoing process, but it can become an object of attention, learning, analysis, communication, and record only to the extent that such processes are apprehended and arrested in presumptively static forms. [24]

Actual occasion features include the following, with emphasized notions given in italics:

- An actual entity is the growing together (*concresce*) of potentials [25]. The process of becoming of an actual occasion is called *concrescence*. The word concrescence is derived from the Latin verb meaning "growing together" [22].
- Actual entities are of a temporal nature, and they come into being and *perish* because of their temporal nature [26].
- "The *enduring* objects of our experience are nothing more than stable patterns of sequential actual occasions" [19; italics added]. Each actual occasion "is a process proceeding from phase to phase, each phase being the real basis from which its successor proceeds toward the completion of the thing in question" [17].
- Actual occasions possess a *subjective* (not conscious) nature that allows them attributes of memory and creativity.
- Complex objects are *societies* (nexuses) of actual occasions that endure cooperatively with emergent unity.
- Actual occasions *prehend* and integrate what the past sends to it by *eternal objects* (patterns/types) [21]. Eternal objects are possible ways in which actual occasions can be definite [25].

It should be pointed out that there is some disagreement among Whitehead scholars as to how far the term *actual entity* can be used to describe that which is commonly held to be an "enduring object" in the contemporary world [7].

B. Actual entities and flow machines

As mentioned, *flowthings* are defined as *what can be* created, released, transferred, processed, and/or received (see Fig. 1). According to Whitehead's ontology, a flowthing is a stable pattern of sequential actual occasions. A flowthing can be visualized as actual occasions that are continuously *becoming*, "each actual entity…is a process proceeding from phase to phase, each phase being the real basis from which its successor proceeds toward the completion of the thing in question" [3].

Fig. 6 gives a general depiction of such a process where, in the context of the source (e.g., a raw material mine), raw materials are created, released, and transferred to the factory, where they are received and processed to trigger the creation of products.

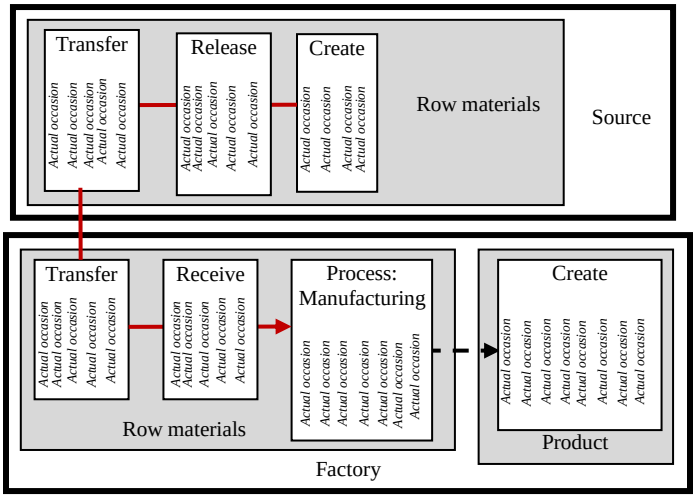


Fig. 6. Flowthings raw materials and product as actual occasions in the context of flow machines

Note the differences in meanings of terminology between FM and Whitehead ontology. The definition of *Process* in Whitehead ontology, given previously, refers to *microscopic* changes in the occurrences of becoming. The *Process stage* in FM is a *macroscopic* phase of the thing that does not create a new flowthing. The *Create* stage in FM refers to the appearance of a new flowthing in the context of a sphere. *Creation* in Whitehead ontology refers to a microscopic *becoming* or emerging into something (actual occasion). It is activity whereby actualities—conceived as individual instances of self-creation—come into being [27-28].

The notion of flow *machine* seems to *bubble up* through some of Whitehead's expressions.

There are . . . two sides to the *machinery* involved in the development of nature. On the one side, there is a given environment with *organisms* adapting themselves to it. . . . The givenness of the environment dominates everything. . . The other side of the evolutionary *machinery*, the neglected side, is expressed by the word creativeness. The *organisms* can create their own environment. [29; italics added]

C. Firehose metaphysics

Bogost [30] identifies a notable weakness in the style of thinking underlying Whitehead's metaphysics:

This is the general sense that for Whitehead reality surges forward like water going through a firehose, one prehension followed by the next without any set of systematic continuities behind, or carrying out, that forward propulsion. Bogost's term for this, "firehose metaphysics," is funny and in some ways apt. [3]

According to Bogost [30],

A process proceeds. First it awakens, then it showers, then it gets dressed, then it brews coffee, then it drives to work, then it opens Microsoft Excel. It travels between two points. Then, then, then, then, then. A metaphysical firehose.

FM presents a different picture of the style, as shown in Fig. 7. The figure depicts *awakening, then showering, then getting dressed, then brewing coffee, then driving to work, then opening Microsoft Excel*. First there is a person (circle 1) in the sleeping state (2), awakening (3) to shower (4), then dressing (5) and brewing coffee (6). Note that for simplicity sake's, the person flow machine is not included in a box. Accordingly, the person goes to his/her car (7–10) to be transported (11) to his/her office (12–15) to open Excel.

This macroscopic description has an interesting variety of processes: Process as movement (from home to office), process as a state (awaken), process as an action (dressing), and process as an agent activity (transporting). An interesting picture, certainly not a firehose, emerges as these variant processes are mixed with triggering, flows, machines, and spheres. Additionally, there are "bricks (actual occasions)" that provide unity and continuity (Fig. 8). More amazing is that these bricks hide the "real" processes inside them. We see here the significance of FM representation in amplifying the true nature of metaphysical description.

IV. Inside the Actual Occasion

At the microscopic level, we can use FM to describe what happens within actual occasion *spheres* (environments), as shown in Fig. 9, showing two instances of actual occasions. Assuming that the actual occasion on the left has already entered the state of *becoming*, or *prehension* (a type of *process* that embeds inheritance in FM; circle 1 in the figure), it is actualized (2) to *perish* while triggering (3) the process of becoming of the next actual (4) that is, in turn, actualized to perish (5), …

In the figure, the process of transitio refers to localizing eternal objects to a space-time region (not shown), and the process of concrescence is the process of coming into being.

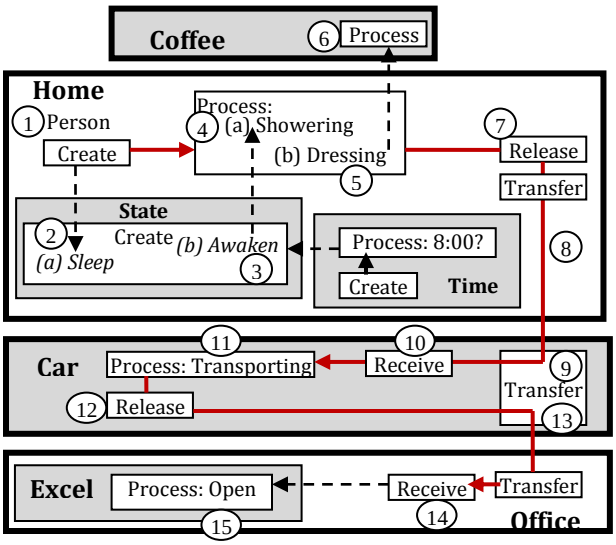


Fig. 7. FM representation of the firehose of processes

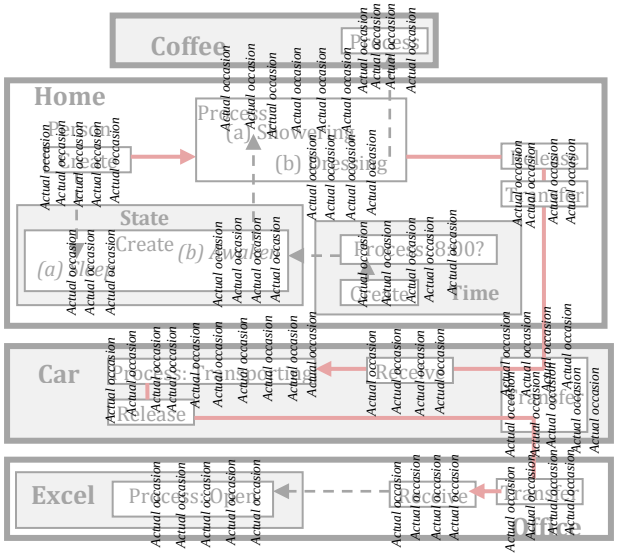


Fig. 8. Unity of the firehose of processes

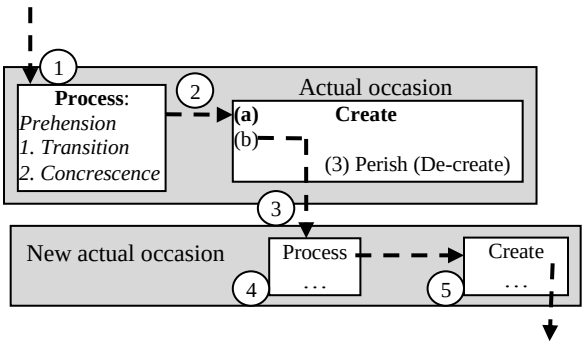


Fig. 9. FM representation of the becoming of a new actual occasion

From the initial set of eternal objects produced by transition, concrescence selects those that are actualized to create the occasion. As mentioned previously (with references),

eternal objects are possible ways (patterns/types) in which actual occasions can be made definite.

Flow (flux) is a change through *prehension*, in the sense of remembering a past (old actual occasion) and anticipating a future (new occasion). We say that the new actual occasion (called the *subject*) prehends the previous actual occasion (called the *datum*). Novelty arises from this prehension; thus "how an actual occasion becomes constitutes what that actual occasion is, so that the two descriptions of the actual occasions are not independent" [3]. Accordingly, prehension involves subject, and novelty (called *subjective form*).

In the literature of process philosophy, the ontological *nature of becoming per se* is a subject of great concern (e.g., [21], [28]). For example, the issues of continuity/discontinuity, unity and diversity, endurance of things, point of completion (satisfaction) of *creation*, duration, succession of two actual occasions,… According to [7], "Whitehead's conception of existence is always focused on the 'how' of becoming (a concrescence of prehensions). For 'how' an actual entity becomes creates what that entity is."

This concern is depicted diagrammatically in terms of the FM *Process* stage: a concrescence of prehensions that triggers a *Create* stage. Thus, schematization in the form of FM representation lends itself to *flowcharting of philosophy* in a systematic way. The result is expression of philosophical thought in computer science language. This merging of the two cultures could be used to establish a more overall view that would further bridge the two disciplines.

It is interesting to study actual occasions in separate macroscopic stages of create, release, transfer, receive, and process. The create stage, at this level, introduces a new flowthing into the system (note that we shift from a philosophical view concerned with existence in nature, to an engineering conceptualization with focus on a part of the world called a system). If this flowthing flows to a process stage, it will experience not only microscopic changes, but also macroscopic change (Fig. 6, previous section). The point here is that the schematization of Whitehead's notion may raise new issues (namely, the effect of different macroscopic FM stages), but here we ignore such observations to pursue the main aim of the paper, which is to introduce this form of representation to facilitate understanding of Whitehead's philosophy.

It is important to note that the purpose of demonstrating FM schemata is to show that this method lends itself to systematic representation of philosophical concepts; thus, some misunderstanding of the real meaning of Whitehead's notion may be reflected. Still, the FM representation acts as a form of language that allows such misunderstanding to be expressed; the diagram can then be redrawn by a philosophy expert to correct the representation if necessary. FM provides a high-level representation of essential concepts and their interrelationships by using diagrammatic notations. Its purpose is to convey a common description without technical specification or written elaboration to facilitate communication between philosophers and nonphilosophers.

V. Example: Rainstorms

Examples of actual (physical) processes include rainstorms, heatwaves, famines, thunderclaps, rumors, performances of symphonies [20]. Consider a rainstorm as a process. An actual occasion of a rainstorm is an instantaneous occurrence, and it's happening is related to other actual entities that overlap one another. This *instance* of a rainstorm is a creative manifesting itself. However, a rainstorm can be conceptualized in FM as a nexus of flow machines, e.g., a rain machine, a lightning machine, wind machine, hail machine, etc. A specific actual occasion of a rainstorm occurs as shown in the upper half of Fig. 10 with fixed rain, wind, lightning, etc. The figure mixes microscopic (actual occasion) and macroscopic (flow machines) views. In the lower part of the figure, another instance of this rainstorm is shown after some change in one or more of its elements, whether rain, wind, or lightning.

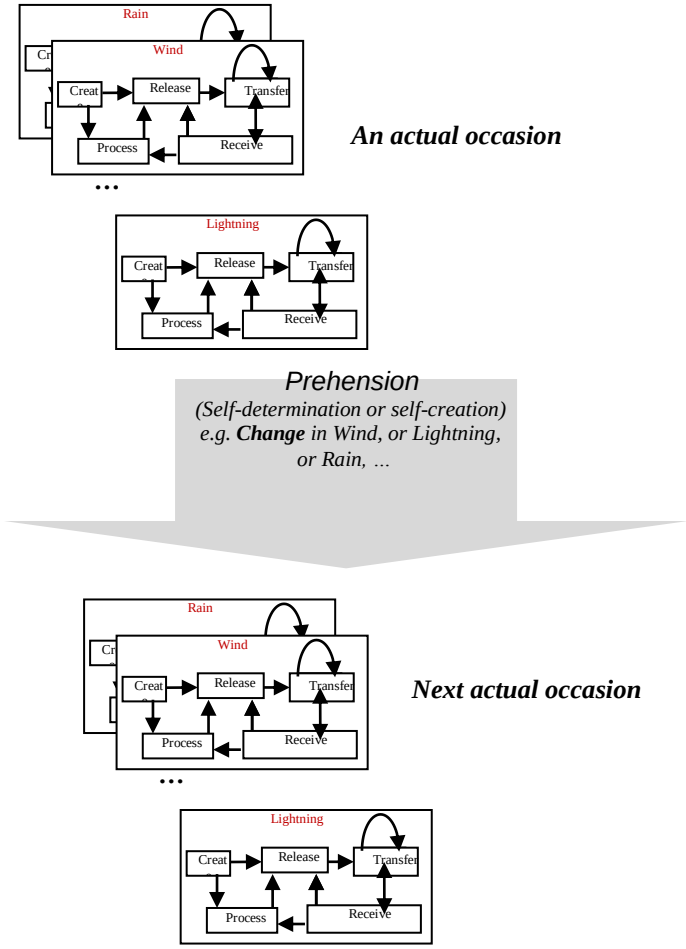


Fig. 10. Rainstorm as a process

As shown in the figure, with *prehension*, the actual occasion of an instance of the rainstorm determines its next instance by an internal change in one of its machines, e.g., now rain is *created*, next the rain is *released* and *transferred* to Earth. Accordingly, the rainstorm is a sequence of rainstorms created again and again, as shown in Fig. 11 (shaded areas in the figure denote earlier occasions). Focusing on *a change*, Fig. 12 shows this process in terms of a change in rain such as becoming heavier.

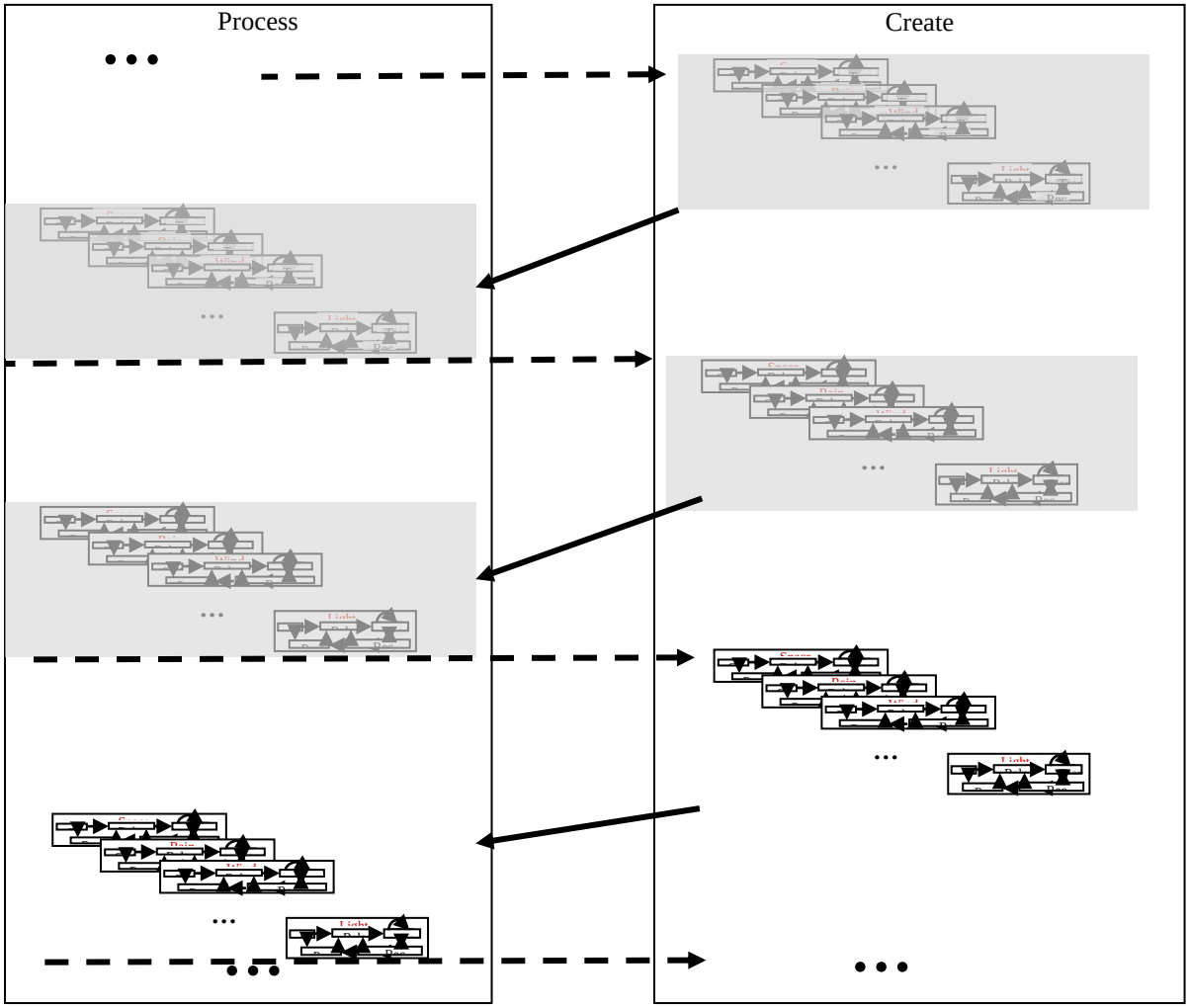


Fig. 11. Sequence of creating instances of a rainstorm

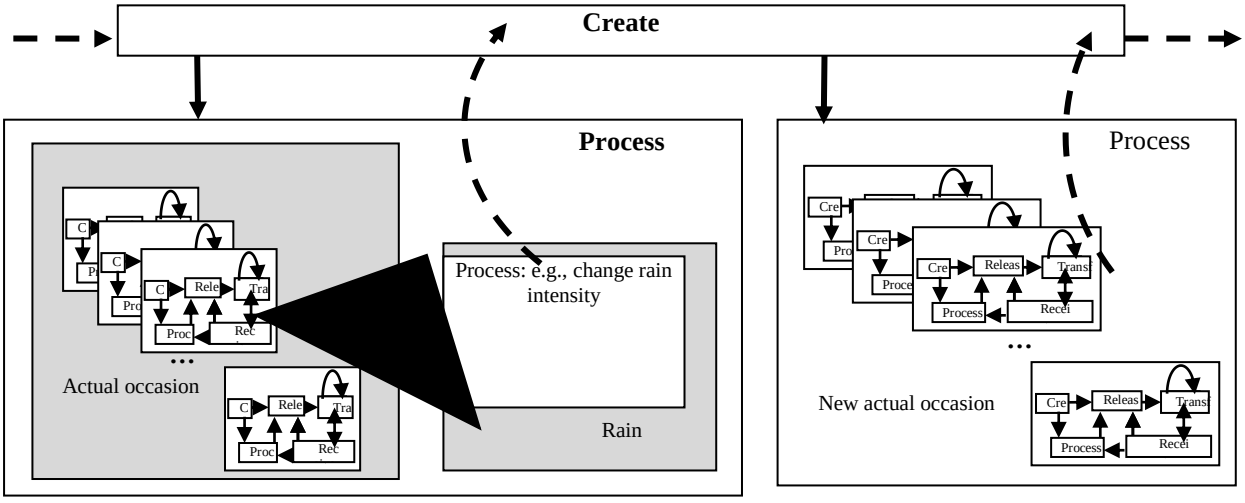


Fig. 12. Example of a change in Rain that causes the creation of a new instance of the rainstorm (dark triangle indicates magnification of a rain machine in the rainstorm)

VI. Cleopatra's Needle

The same type of representation can be applied by treating a solid and permanent object as an *event*. Whitehead [2] gives the example of Cleopatra's Needle, an obelisk situated on the Victoria Embankment in London. For Whitehead,

Cleopatra's Needle isn't just a solid, impassive object upon which certain grand historical events [actual changes]—being sculpted, being moved—have occasionally super-vened. Rather, it is eventful at every moment. From second to second, even as it stands seemingly motionless, Cleopatra's Needle is actively *happening*… At every instant, the mere standing-in-place of Cleopatra's Needle is an event: a renewal, a novelty, a fresh creation. That is what Whitehead means, when he says that events—which he also calls "actual entities" or "actual occasions"—are the ultimate components of reality. [32]

A physicist who looks on that part of the life of nature as a dance of electrons, will tell you that daily it has lost some molecules and gained others, and even the plain man can see that it gets dirtier and is occasionally washed. [3]

Fig. 13 shows two consecutive instances of Cleopatra's Needle. The flow machines of electrons and dirt (exemplified in the above quote) are drawn as *complete* flow machines to indicate that any change can happen, e.g., receipt or output of electrons… "Cleopatra's Needle is a society [the grouping of actual occasions], or an enduring object" [32].

We interpret an event (actual occasion) in terms of changes as shown in the figure. The following quotes from Stoney [33; italics added] shed some light on the process of becoming,

- "Events have some capacity, however slight, to select among alternatives" (circle 1 in the figure).
- "Each event feels the feelings of – is connected to [*prehension*] – earlier events." (circle 2). Feeling, here, does not refer to a conscious experience.
- "Events have aims (*goals*; circle 3), namely to maximize creativity and intensity of feeling, that arise due to their participation with more dominant occasions of experience."
- "This process of self-determination is *concrescence*… existence is a series of coming into beings".
- "An event that has completed its concrescence has achieved *satisfaction.*"
- "The dominant occasion of experience [enduring objects - patterns] integrates the lower level actual occasions into a unity of purpose. For human beings, the dominant occasion of experience constitutes the mind."
- "For any actual occasion, the future is open, i.e., unpredictable because of the alternatives available to it. This is the basis for the appearance of novelty."

Such notions as prehension (synonym: feeling, as used by Whitehead) and goals can be incorporated into the diagram. They are machines, just like the physical machines of rain, wind, lightning, and hail. Actual entities follow each other like drops of experience. An instance of a new occasion becoming occurs with prehension, i.e., connection to an earlier actual occasion [33]. Actual occasions begin; live their lives, attain completion (satisfaction), and perish [25].

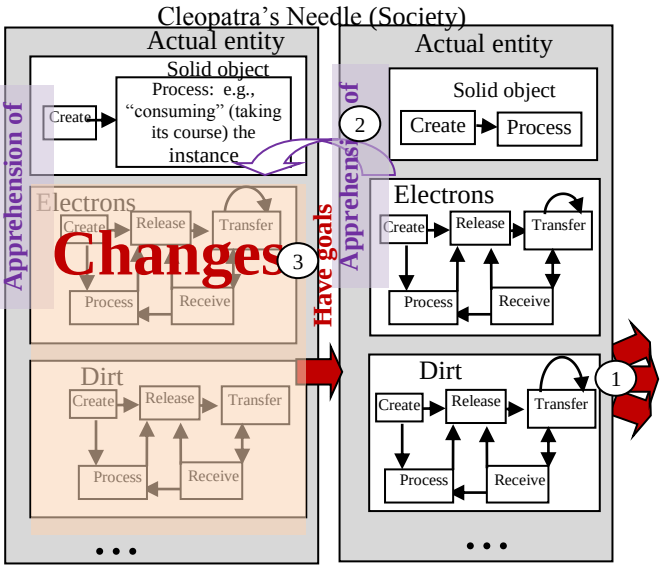


Fig. 13. Cleopatra's Needle as a process

VII. Conclusion

This paper has attempted to employ schematization *to understand* philosophical concepts. While the method is applicable to several philosophical works, it focuses specifically on a portion of Whitehead's ontology that is based on the notion of process. The approach uses a diagrammatic modeling tool to produce a conceptual representation of such notions as actual occasions, process, becoming, becomingness, and actual in contrast to potential. The resulting representation seems to introduce a new method of discussion of meanings embedded in Whitehead's philosophy. This initial attempt points to its viability in this context and is worthy of pursuit.

The paper hints at examining the notion of becoming in different macroscopic stages of the states of creation, release, transfer, receiving, and processing. That is, do these states of a *thing* have an effect on the relevant Whitehead processes?

References

[1] A. Schwill, "Fundamental ideas of computer science," Bull. Eur. Assoc. Theor. Comput. Sci., vol. 53, 1994.

[2] P. Rose, On Whitehead. Belmont, CA: Wadsworth, 2002.

[3] A. N. Whitehead, The Concept of Nature. Prometheus Books, 2004 (originally published 1920). www.prometheusbooks.com/

[4] J. Palomäki and H. Keto, "A process-ontological model for software engineering," in CAISE'06, 18th Conference on Advanced Information Systems Engineering—Trusted Information Systems, Luxembourg, T. Latour and M. Petit, Eds. Proceedings of the Workshops and Doctoral Consortium, pp. 720-726.

[5] I. Stengers, "A constructivist reading of Whitehead's Process and Reality," Theory Cult. Soc., vol. 25, no. 4, pp. 95–96, 2008.

[6] L. Armour, "Looking for Whitehead," Br. J. Hist. Philos., vol. 18, no. 5, pp. 925-939, 2010. DOI: 10.1080/09608788.2010.524768

[7] M. Halewood and M. Michael, "Being a sociologist and becoming a Whiteheadian: toward a concrescent methodology," Theory, Cult. Soc., vol. 25 , no. 4, pp. 31-56, 2008.

[8] T. E. Eastman and H. Keeton, Eds., Resource Guide for Physics and Whitehead, supplement to Physics and Whitehead: Process, Quantum and Experience. Albany: State University of NY Press, 2003.

[9] S. Krämer, "Epistemology of the line. Reflections on the diagrammatical mind," in Studies in Diagrammatology and Diagram Praxis, A. Gerner and O. Pombo, Eds. London: College Publications, 2010, pp. 13–38.

[10] S–J. Shin, O. Lemon, and J. Mumma, "Diagrams," The Stanford Encyclopedia of Philosophy, Winter 2014 edition, Edward N. Zalta, Ed. http://plato.stanford.edu/archives/win2014/entries/diagrams/.

[11] G. Dodig-Crnkovic, "Shifting the paradigm of philosophy of science: philosophy of information and a new renaissance," Minds Mach., vol. 13, pp. 521–536, 2003.

[12] R. Turner and A. H. Eden, "The philosophy of computer science," J. Appl. Logic, vol. 6, no. 4, p. 459, 2008.

[13] I. Ventura, On Philosophy and Software Design. http://www.philosoftware.com/about [accessed October 16, 2015]

[14] S. Al-Fedaghi, "Alternative representation of aspects," 10th IEEE International Conference on Information Technology: New Generations, IEEE ITNG 2013, Las Vegas, USA, 2013.

[15] S. Al-Fedaghi, "Flow-based specification of time design requirements," Int. J. Adv. Comput. Sci. Appl. (IJACSA), vol. 6, no. 8, 2015.

[16] S. Al-Fedaghi, "Diagrammatic representation as a tool in clarifying logical arguments," Int. J. Adv. Res. Artif. Intell. (IJARAI), vol. 4, no. 10, 2015.

[17] A. N. Whitehead, Process and Reality: An Essay in Cosmology, Corrected Edition, D. R. Griffin and Donald W. Sherburne, Eds. New York: The Free Press, 1978.

[18] Learner.org, Mystery Class, "Revolution of the Earth," accessed 2015. https://www.learner.org/jnorth/tm/mclass/Glossary_rev.html

[19] R. L. Stein, "Towards a process philosophy of chemistry," Int. J. Philos. Chem., vol. 10, no. 1, pp. 5-22, 2004.

[20] N. Rescher, Process Metaphysics: An Introduction to Process Philosophy. New York: State University of New York Press, 1996.

[21] S. Rosenthal, "Continuity, contingency, and time: the divergent intuitions of Whitehead and pragmatism," Trans. Charles S. Peirce Soc., vol. 32, no. 4, 1996.

[22] D. B. H. Farr, "A critical examination of A. N. Whitehead's metaphysics in light of the later Martin Heidegger's critique of onto-theology," PhD diss., McMaster University, Ontario, Canada, 2005.

[23] J. B. Cobb Jr. and D. R. Griffin, Process Theology: An Introductory Exposition. Philadelphia, PA: Westminster Press, 1976.

[24] R. S. Perinbanayagam, Identity's Moments: The Self in Action and Interaction. Lexington Books, 2012.

[25] D. A. Crocker, "A Whiteheadian theory of intentions and actions," PhD dissertation, Yale University, 1970.

[26] M. T. Teixeira, "The stream of consciousness and the epochal theory of time," Eur. J. Pragmat. Am. Philos., vol. 3, no. 1, 2011.

[27] L. Ford, "On epochal becoming: Rosenthal on Whitehead," Trans. Charles S. Peirce Soc., vol. 33, no. 4, pp. 973–979, 1997.

[28] L. Ford, The Emergence of Whitehead's Metaphysics, 1925-1929, SUNY Series in Philosophy , June 1985. ISBN10: 0-87395-857-8

[29] A. N. Whitehead, Science and the Modern World. New York: Free Press, 1967, pp. 111–112.

[30] Bogost, "Process vs procedure" in Fourth International Conference of the Whitehead Research Project, 2011.

[31] A. J. Ivakhiv, The attractions of process (metaphysics) [blog post], December 9, 2010. Accessed Oct. 20, 2015. http://blog.uvm.edu/aivakhiv/2010/12/09/the-attractions-of-process-metaphysics/

[32] S. Shaviro, Without Criteria: Kant, Whitehead, Deleuze, and Aesthetics. Cambridge, MA: The MIT Press, 2009.

[33] S. D. Stoney, "Towards an ecological neuroscience: aspects of the nature of things according to process philosophy," 2001. http://home.earthlink.net/~icedneuron/ProcessPhilosophy.htm

Mood Extraction Using Facial Features to Improve Learning Curves of Students in E-Learning Systems

Abdulkareem Al-Alwani
Computer Science & Engineering Department
Yanbu University College, Royal Commission Institute & Colleges
Yanbu, Saudi Arabia

Abstract*—Students' interest and involvement during class lectures is imperative for grasping concepts and significantly improves academic performance of the students. Direct supervision of lectures by instructors is the main reason behind student attentiveness in class. Still, there is sufficient percentage of students who even under direct supervision tend to lose concentration. Considering the e-learning environment, this problem is aggravated due to absence of any human supervision. This calls for an approach to assess and identify lapses of attention by a student in an e-learning session. This study is carried out to improve student's involvement in e-learning platforms by using their facial feature to extract mood patterns. Analyzing themoods based on emotional states of a student during an online lecture can provide interesting results which can be readily used to improvethe efficacy of content delivery in an e-learning platform. A survey is carried out among instructors involved in e-learning to identify most probable facial features that represent the facial expressions or mood patterns of a student. A neural network approach is used to train the system using facial feature sets to predict specific facial expressions. Moreover, a data association based algorithm specifically for extracting information on emotional states by correlating multiple sets of facial features is also proposed. This framework showed promising results in inciting student's interest by varying the content being delivered.Different combinations of inter-related facial expressions for specific time frames were used to estimate mood patterns and subsequently level of involvement of a student in an e-learning environment.The results achieved during the course of research showed that mood patterns of a student provide a good correlation with his interest or involvement during online lectures and can be used to vary the content to improve students' involvement in the e-learning system.More facial expressions and mood categories can be included to diversify the application of the proposed method.

Keywords*—Mood extraction; Facial features; Facial recognition; Online education; E-Learning; Attention state; Learning styles

I. INTRODUCTION

The main problem arises in E-Learning as there is no supervisor to assess how students are physically and emotionally responding to the delivered content. Usually when the students taking any course online, they may lose concentration and focus resulting in poor academic performance. Tackling this issue can advance the e-learning process many fold as each student's interest can be assessed and necessary improvements can be made to the content to engage the user during the online lecture. In order to circumvent the problem of observing student on an e-learning platform, this research is conducted with a view to analyze the relationship between facial expressions of a student enrolled in an e-learning system and the ways to improve upon learning attitude of such students using information extracted from these features.

E-learning is a medium for imparting education anytime and anywhere, and due to recent advances in information technology, online education systems can be considered as a blessing and an important information technology asset. Knowledge transfer via informational technology tools requires careful planning and execution as the learning environment provided to the student during e-learning offers complex insight on the student's learning curve. In order to improve the e-learning experience, the process of learning becomes imperative as it majorly governs how much and how well a student can absorb knowledge during online lectures [1].Delivery of content, examinations and student feedback are important measures that have a direct effect on the learning curve of students as well as the e-learning objectives. Still the time frame required for relating and observing all these measures must be long enough to account for every possible detail [2].

These measures are also the same as that are used in traditional or on campus learning where teacher has a direct interaction with students. Initially computers and information technology was used as tools to improvise learning. This concept subsequently evolved to full-fledged e-learning systems. Universities have now started offering online courses and have developed e-learning platforms catering to the need of almost any student. E-learning has allowed off campus students to get educated at homes or simply anywhere in the world.

Knowledge delivery through e-learning offers numerous advantages but most of its features can only be fully utilized if the student's involvement and interest remains continuous throughout the course of online education [3].As a student has a personal preference for acquiring knowledge at one's own time and pace, this allows people from all walks of life to have an opportunity to learn and educate themselves without any restrictions of time and space.

With this evolution in the e-learning technologies and the increasing number of students, requirements for improving online education experience are getting more and more demanding. It's understood that more in depth studies are

needed in order to ascertain the variables which can really affect online educational environment in a positive way [4].

Natural feedback on the content being delivered can be taken automatically from learners by using their facial expressions as a tool to measure interestingness of the content and engagement of student in the online lecture [5].Facial expressions can provide critical information on student's interest and participation in online educational learning. Faces provide detailed information about an individual's state of mind, mood and also emotional state. Studies throughout history have shown that facial expressions are the prime representation of human emotions. Facial expressions can be considered as the main source of information, after words, in estimating an individual's thoughts and state of mind [6].

Facial Recognition has proven to be an important tool in automate tutoring as it helps in the improvement of students learning outcomes as well as in the development of the learning experience [7][8]. In the end, this leads to improvement in the learner's involvement in the learning environment.

This research aims at enhancing students' learning outcomes while studying online courses.This can be considered as analyzing the real-time interaction between student and machine, and assessing student's engagement during E-learning session, which is constantly changing over the passage of time. This variable of engagement can be plotted against time and can be considered as a function of time. This engagement function will be called a student's learning curve in the rest of this paper as its variation as a function of time directly affects the learning aptitude and interest of the students.

The basic claim made in this study is that lack of students' involvement/engagement during online classes due to the lack of the physical presence of teachers is the main factor that hinders learners from achieving on-line courses' learning outcomes. This is largely due to the absence of any direct teacher supervision of the students learning process who in such learning context may be distracted in many ways from what they are studying ,with there is no one present to supervise them in what they are learning.

A student studying using online resources cannot participate in a verbal communication, then the major attributes that can be observed to ascertain a student's mood and attitude are his facial features and body language [9].

The prime motive behind this study was to devise a methodology to identify major mood patterns with high probability in an e-learning environment. Data continuously pile up when visual data is recorded in real-time. The sample space becomes large and takes more computational power. To address this specific issue, the secondary goal of this research was to integrate a sequential mining technique which can identify mood patterns with high probability. Rules were extracted using Apriori algorithm to reduce the mood sample space by tagging frequent facial feature patterns into predefined five mood categories.

In the subsequent sections, literature is reviewed followed by a discussion of research methodology for applying the proposed technique, and in the end the results are presented with concluding remarks.

II. Literature Review

E-learning presents a lot of learning opportunities for people unable to attend regular schools, colleges or universities. Given the importance of E-learning in this information age, a lot of research has been carried out to improve the performance and adaptability of e-learning. This section will present past, present and prospective studies undertaken for the purpose of improving the e-learning eco-system.

Online teaching and e-learning methodologies have transcended to new levels after the boom of information technology age. As a result, the quality of education and number of online learners has increased substantially. Still, the modernized way of e-learning creates problem that affects a student's learning curve due to unavailability of any direct supervision [10].

An instructor can provide some insight into student's satisfaction during lectures [11], therefore student's involvement in class has direct correlation with the professional aptitude of the instructor[9]. Direct supervision not only facilitates learning but also keeps the student synchronized with the course objectives due to instant communication with the instructor at any time during the lecture. Lack of communication has shown that affected students may experience high levels of frustration [11].

As supervised teaching is very critical to the learning curves of the students, online courses present a different set of challenges to instructors and students. Online students may never visit a physical campus location and may have difficulty establishing relationships with faculty and fellow students. Researchers who study distance learners must understand and account for these differences when investigating student satisfaction [12],mentioned three important types of interaction in online learning courses: (a) learner-content,(b) learner-instructor, and (c) learner-learner. He emphasized that instructors should facilitate all types of interactions prompting attentiveness in their online courses as much as possible.

E-learning requires use of video, audio, text to simulate the traditional class and learning environment as closely as possible. E-learning environments may be used for a numerous educational purposes. Modern trends indicate that e-learning based education will come at par with traditional education methods in the near future. In an e-learning environment, teacher and student are not in direct interaction and content is provided by the instructor thorough online platforms using multimedia and software interfaces.

As there is no means of instant communication, machine can only understand what it records using standard man machine interfaces. As there is no verbal communication between the student and the e-learning platform, facial expressions are the only means that can provide concrete information about a student's mood and involvement during the class [13]. For example, when students show confused expressions, one of the common mood patterns may be one or a combination of the following facial features i.e. eyebrows

lowered or drawn together, vertical or horizontal wrinkles on the forehead, and inconsistent eye contact etc. In order to understand whether the student is grasping what is being delivered, a lecturer must sense the subtle nonverbal indicators exhibited by the expressions of the students [14].

Facial features and there relevance to emotions has been rigorously investigated by Ekman et. al [26][27][28][29]in various publications and their work is regarded as one of the most significant contribution to facial attributes based emotion analysis. Facial acting coding system can provide information about instantaneous facial emotional reactions, but still the need to ascertain a complete mood based on various action units as they vary from person to person and situation to situation.

Facial features (Forehead, eyes, nose, mouth, etc.) are the fundamental attributes that are extensively used in face recognition systems as their movements help determine the construction of expression on a human face [15].

Facial recognition can be efficiently used to identify and categorize facial expressions in real-time. Machine learning algorithms have also been employed for facial recognition to enhance accuracy and detection time[16]. Facial expressions are basically emotional impulses translated into physical muscle movements such as, wrinkling the forehead, raising eyebrows or curling of lips. Authors in [17] presented the beneficial prospects of using intelligent methods to extract facial expressions to improve the processing speed of image analysis. Database of facial expressions have been populated in various studies to develop interesting algorithms for various applications.

Emotion recognition study can be broadly categorized into three steps: Face detection, Facial feature extraction and Emotion classification. Detailed research has been carried out in each of these. These three categories are concerned with the central background pertaining to the issue of facial emotion recognition.

In an image, detecting the presence of a human face is a complex task due to the possible differences attributed to different faces. The varying physical attributes of a face are the major cause for this variation. The emotions which are the combination of facial action units [31] in a human face also affect facial appearances.

Neural networks can be actively used to classify a learner's orientation in predetermined categories, which can be associated using Apriori algorithm to allow for real-time HMI intervention for improved involvement. The aim was to assess in real-time whether the e-learning systems can be improvised to recognize the facial expressions and attention state of a learner using classification and data association algorithms. These systems can then be used to improve content delivery of e-learning platforms through real-time mood extraction. Appropriate learning materials and activities for a learner can be incorporated to alter his mood state during e-learning activity.

Face detection can be broadly classified into four categories: knowledge-based approach, feature invariant approach, template-based approach and appearance-based approach [33][34][35][36].

Appearance-based approach maps the human face in terms of a pixel intensities. Since only face patterns are used in its training process, the efficiency is not good. Even the time taken is lengthy, as the number of patterns which needs to be tested is large.

A neural network was found to be quite effective in capturing complex facial patterns from facial images. Both supervised and unsupervised learning approaches are used to train the neural network. Since finding a sufficient training data set is questionable, unsupervised neural networks are more preferable. Apart from neural networks, Support Vector Machines (SVM)[37], eigenfaces, Distribution based approaches, Nave Bayes classifiers, Hidden Markov Models (HMM)[38] and Information theoretical approaches can also be used for face detection in the appearance-based approach [33][34][35][36]. Rather than minimizing the training error as in neural networks, SVM operate by minimizing the upper bound on the generalization error limit instead of minimizing training error as in neural networks. Eigen faces uses Eigen space decomposition and has proven an accurate visual learning method. Nave Bayes classifier is more efficient in estimating the conditional density functions in facial sub-regions. The HMM differs from template-based and appearance-based approaches as it does not require exact alignment used in these approached rather HMM constitutes a face pattern as a series of observation vectors.

A student involvement in e-learning is directly based on how he can be engaged to focus and listen to the content being delivered. Facial expressions over short instants can be misleading and a time frame based analysis to ascertain emotional states can provide interesting results. For example, confusion and frustration was studied using temporal and order based patterns using continuous affect data [30]. A similar study was carried out by Craig et al [32] which also included boredom. Authors reported that confusion is affiliated with indirect tutor dialogue moves and negative tutor feedback. Similarly, frustration was found to be affiliated with negative tutor feedback, and boredom did not appear to be detectable from the set of three dialogue features [32]. Timing of an emotional state can also play an important role in automated tutoring as reported by authors in [31].This study investigated the relationship between affect and learning. However, identifying the exact places where emotion occurred during the learning process was not covered limiting the efficacy of Auto Tutor system.

III. RESEARCH OBJECTIVES

A coherent information exchange between learner and machine is imperative for effective E-learning and is based on the learning curve of the student. Research objectives in this study were formulated to develop a practical technique for understanding student interest during the E-learning session.A student interest can thus be enhanced vide engagement techniques. The research objectives for achieving this objective are listed as follows:

First objective was to investigate whether facial expressions are the most pertinent means of nonverbal expression mode during e-learning and can in turn assist the e-learning system to identify the interest and comprehension level of the students.

Second objective was to list most common facial features that describe the involvement of a student in a lecture. A list of 54 features were compiled and used in a survey to identify most pertinent facial features for describing student's expression.

The third objective was to develop a methodology to relate facial features to understand expressions of a student during various emotional states describing his involvement in the lecture with high probability in real-time.

Next section consists of the methodology pursued to identify important facial features and will present details on how facial features recorded over certain time frames can provide sufficient information regarding moods of a student in real-time with reduce computational power and sample space.

IV. Research Methodology

Research methodology pursued in this research was conducted phase wise. First, a survey was carried out among instructors involved in e-learning to inquire and identify most probable facial features that represent the facial expressions, and over certain time, the mood patterns of a student. A neural network approach is then used to train the system using facial feature sets to predict specific facial expressions. Data association based algorithm was selected in proposed approach to extract information on emotional states by correlating multiple sets of facial features using support and confidence levels. This was done to improve the clustering of the relevant datasets. The methodology was designed to analyze a student's interest by varying the content being delivered.Different combinations of inter-related facial expressions for specific time frames were used to estimate mood patterns and subsequently level of involvement of a student in an e-learning environment.

A trained dataset of facial features representing student's emotional state is the primary requirement to assess a student's involvement in an e-learning environment. The data has to be collected first, correlated with emotional indicators and is then reused as training data to extract different expressions describing facial features. The data association algorithm is applied to relate features into expression over a time period to discover negative mood patterns of a student during online lecture. The methodology presented here in order to pursue above objectives consists of three major phases, which are as follows:

A. Categorization of Facial Features Using a Survey Instrument:

Before embarking on a detailed investigation into efficacy of observing facial features to asses a student's interest during an online lecture. A survey is conducted to evaluate whether facial feature analysis is the most pertinent means of understanding a student's behavior during e-learning. Secondly, the survey recorded observations from academics regarding which facial features partially describe mood or emotional state of a student.

In order to construct a baseline for the facial features, a survey was conducted in 2014/2015 academic year and 198 academics from various universities were approached for their response. Experts in online instruction were approached with minimal 2 years of teaching experience at postsecondary level.

Survey was forwarded with a brief explanation of the research objectives with two instrument questions which were,

- Will the process of measuring the learners' degree of engagement/involvement during studying online courses help in learner to focus more and as a result improve the learning outcomes? Which facial expressions you think are most obvious and recurrent in lectures?
- List the most pertinent facial features for eyes, eyebrows, lips and head that constitute facial expressions of student, representing his state of mind and involvement during a lecture.

B. Training and Classification of Facial Features Dataset using Neural Networks:

Classification algorithms make use of supervised learning techniques to predict the class of previously unobserved data by using a training model from existing data [18]. An efficient way to define a classification model is to characterize it as a set of comprehensive classification rules to provide relevancy and accuracy, simultaneously. An extensive 'Cohn-Kanade' data was selected for training and classification of our neural network model.

The Cohn-Kanade AU-Coded Facial Expression Database [19] is available for research purposes online and is used in facial image analysis and for perceptual studies. This database consists of 486 sequences from 97 faces. Each sequence starts with a neutral expression, gradually leading to the peak

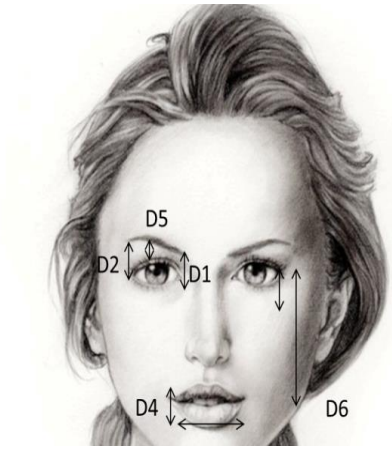


Fig. 1. Distance attributes helping in measuring facial features

expression. This database provides solid foundation for our trained NN model, speeding up the face recognition process.

Feature collection can be carried out using feature extraction algorithms presented in literature. Recognition and

interpretation mood or learning attitude of a student is carried out by analyzing facial features during the lecture. Neural networks are used to train our system on 'Cohn-Kanade' facial expression database and same was used to identify a student's involvement state during an online lecture. Facial features were characterized into four main categories based on the response of the survey: eyes, eye brows, lips, head (incl.hand/fingers on face).

A Radial basis function NN algorithm [20]which was used in this study to classify facial expressions based on facial features(Table-1). Figure-1 shows the distance points for eyes, eyebrows and lips that were used to define facial features for the training of NN algorithm. Any change in the distance metrics point to a certain facial feature instance and combination of these facial features can be used to classify the five facial expressions.

As distance provides certain thresholds to make decisions related to facial expressions, these thresholds can be used to classify unknown patterns [25]. Figure-2 shows the higher level decision model of NN used in our proposed model for classification of facial features.

In this figure the pixels and combinations of two distance attributes are shown which provide information on eyebrow position relative to the eye. The matrix shown here can be used to plot a feature if D1 and D2 are plotted over a 2-D plot. The proposed classification model is already trained to identify the affinity of the region belonging to the acquired D1 and D2 values, and it correlates the D1 and D2 values to a certain feature, i.e., in this case Feature '21' based on the prior training data. Therefore it is necessary to train the proposed model with a larger database to improve the probability of detection of a certain feature.

C. Mood Extraction Using Facial Features

Associative Data Mining is considered as an important data mining technique and it has been extensively researched and used for data mining by researchers. Data Association helps in mining of association based rules between items based on item set transactions and it is regarded as an important tool for rule discovery in very large datasets [21]. Data association can provide an estimate of unknown relationships and decision rules in a dataset which can greatly improve the process of decision making and prediction [22].

A student's mindset can be well communicated through his facial expressions during an online lecture. Change of mood of a student can be observed using following instruments: facial expressions, hands and body language. These instruments can be observed individually or in combination, however, in both the cases, the data association patterns can be extracted to get a better understanding of the behavior of the student during online learning. This data association approach is very efficient and provides accurate result in instances when one category alone cannot be used to assess accurate understanding of the state of mind of a student [24]. A combination of facial feature categories which have a high probability of occurrence reduces the decision space by many folds. Facial expressions for each student for every online course can amount to very large data. Therefore, a well-established algorithm is presented here to extract moods from a large set of facial features. The algorithm is modified to be used for identifying moods of students and subsequently making accurate decisions about their interest level during the delivery of an online course. Mood extraction using data association is carried out in two stages [24] [25].

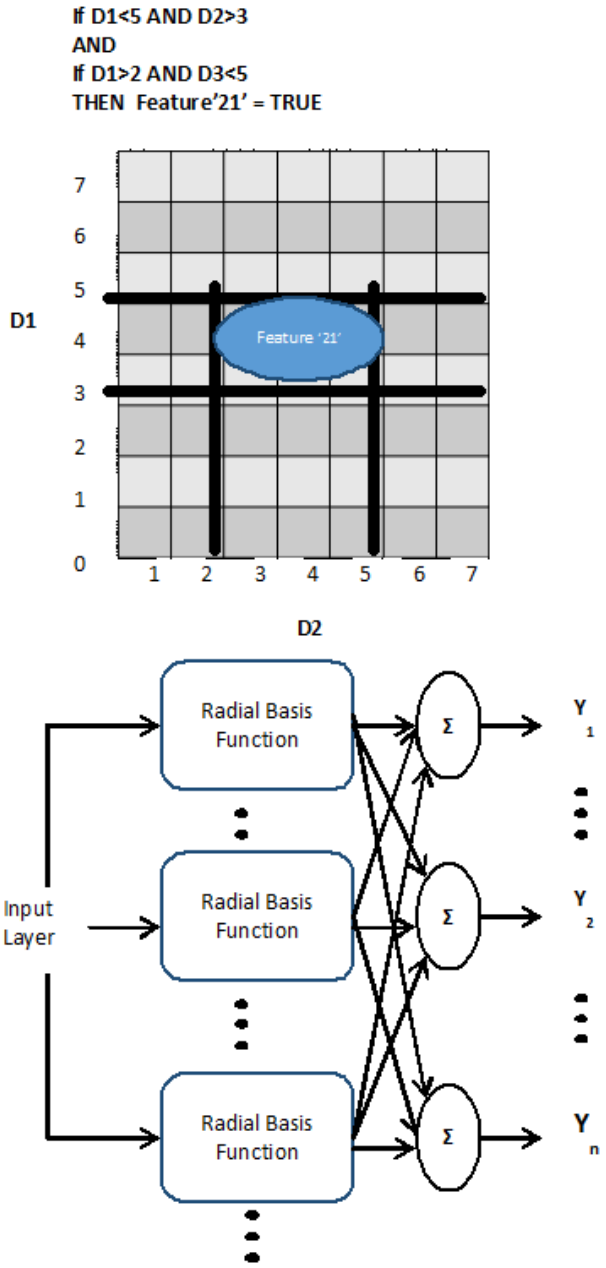


Fig. 2. A distance based radial basis function decision approach for facial features

Approximating categories or item-sets, that occur frequently, and data association for extracting rules that are based on the relationships between these items is the first step of classification. In the subsequent stage,, items are evaluated to segregate sets of items that occur frequently and have a ratio of occurrence greater than the minimum support threshold [23].

In the second phase, all possible rules are extracted from the item set, and the number of rules will depend upon the all possible combinations of the items in a given item set e.g. if an item set is of the form {a1, a2, a3}, then the rules that can be extracted are {a1→a2, a3},{ a2→a1, a3}, { a3→a1, a2}, { a1 , a2→ a3}, { a1 , a3→ a2}, { a3 , a2→ a1} etc.

A rule {X→Y}, where X and Y are facial features can be verified using confidence and support threshold levels. Support and confidence thresholds are used as constraints or limits for rule extraction. Support and confidence thresholds provide a measure for pruning the rules which does not meet the threshold criterion. In a nutshell, associative data mining is used for mood extraction by employing user specified support and confidence levels for related facial features and this approach can be used to devise a gauge forassessing extent of correlation between facial features in a dataset.

Apriori algorithm [24] is an efficient technique for data association which can be used to generate frequent feature sets from database facial features. The algorithm [25] makes iterative estimation of most frequent items based on support and confidence metrics. Other metrics may also be used but these are the standard metrics in assessing frequency of an item set and objectiveness of a student's mood during online learning.

Support level is used to estimate that how often a relationship is established in between various facial features in a dataset, while confidence level provides a measure to determine the frequency of 'B' facial feature in observed features also containing feature 'A' during time 't'. Time period for observing a student's involvement and attentiveness is dependent upon the content being delivered and significant parts from the content that require complete attention from the students for understandability.

Support determines how frequent is a distance attribute appearing in SET 'A' also appear in SET 'B' for a given number of samples, whereas, confidence determines how frequently distance attributes from SET 'B' correlates with distance attributes in SET 'B'.

Abovementioned support and confidence metrics can be mathematically represented as

Support:

$$s\,(A \rightarrow B) = \sigma(A \cup B)/N \qquad (1)$$

Confidence:

$$c\,(A \rightarrow B) = \sigma(A \cup B)/\sigma(A) \qquad (2)$$

Following steps are proposed for robust mood extraction architecture to evaluate interest and attention of a student towards educational content being delivered online. These steps will be form the basis of the proposed algorithm for mood extraction using facial features.

- Frequent Feature-sets Generation: Considering N transactions, all frequent feature-sets are estimated based on support levels. This is an iterative process to identify and generate candidate feature-sets. This part of the algorithm involves two phases. In first phase, it checks each feature-set starting from single facial feature from the feature-set to the maximum size feature-set. In the second phase, new feature-sets are estimated from the previous iteration and support is tested against the support threshold. Number of iterations in this step depends upon the maximum size of item set i.e. (kmax + 1) is the total number of iterations and kmax is the largest size of a frequently occurring feature-set.
- Candidate Generation and Pruning: In this step, new candidate feature-sets are generated based on the (k-1) feature-sets found in the previous iteration followed by pruning by using support levels.
- Support Counting: In this step, occurrence frequency of candidate feature-sets after pruning is determined and support levels are updated.
- Mood Extraction: A level-wise approach is used to discover rules based on data association between consequent and antecedent facial features in frequent feature-sets. At first, all the rules with a single consequent are selected to generate new candidate rules. The selection of these rules is based on respective confidence levels. Rules generated by Apriori algorithm can be large in number depending upon the database being searched.

As a test case, 30 students from a mathematics class were observed during a one hour session of e-learning and expressions were extracted using Apriori method explained above. Lecture session was divided into 10 minutes sub-sessions, where each sub-session addressed a particular mathematics problem. Table-1lists the 10 minute divisions of a one hour lecture.

TABLE I. MODULES TAUGHT DURING 6 X 10 MINUTE TIME FRAME IN ONE HOUR SESSION

S.No	Sub-session
1	Introduction
2	Matrices
3	Matrices Multiplication
4	Matrices Division
5	2 x 2 Matrix operations
6	4 x 4 Matrix operations

Students with an average age of 15 years were selected for the study. All students were from grade 10 in a private school. Students were selected based on their academic performance and sufficient exposure to e-learning environments. No prerequisite information was provided to them regarding nature of this exercise.

A 35mm digital camera was used with a 10 fps frame rate to record facial features. Using association between facial features, facial expressions were sought out to extract mood of a student for 6 x 10 minutes time frame during learning. Radial basis function based NN algorithm [20] was employed to classify moods based on facial features. Apriori algorithm is subsequently used to create frequent feature sets or mood sets from which most pertinent rules can be extracted to declare a mood pattern valid.

A written feedback was acquired from every student after each 10 minutes session comprising of the following two questions:

1) Mention in which parts of the lecture you were

- Happy
- Sad
- Confused
- Disturbed
- Surprised

2) Which parts of the 10 minute frame did you not understand or were inattentive (1 to 10)?

Based on this topology all three phases were executed sequentially, and results for the study are presented in the next section

102 out of 200 participants provided their feedback based on which result of the survey were formulated and was used to categorize five major expressions and 23 facial features. These are listed in Table-2 and Table-3.

An astounding 88 percent of the respondents agreed that facial expressions do reveal the involvement of a student in the class and can be used to assess a student's response to the content being delivered. This provided strong basis for our subsequent analysis, which was carried out to classify facial features using neural networks.

TABLE II. FACIAL EXPRESSIONS DEFINING MOOD OF A STUDENT IN A CLASSROOM

S.No	Facial Expressions	Freq.
1.	Happy	91
2.	Sad	96
3.	Confused	94
4.	Disturbed	90
5.	Surprised	87

TABLE III. MAJOR FACIAL FEATURES THAT CONSTITUTE THE FIVE MAJOR FACIAL EXPRESSIONS OF A STUDENT

S.No	Facial Features	Freq.
1.	Eyes focused on the screen	90
2.	Eyes enlarge	75
3.	Eyes shrink	76
4.	Eyes rolling	81
5.	Eyes blinking	92
6.	Eyeshot in contact with screen	85
7.	Wide-open eyes	80
8.	Lips tight	67
9.	Lips reading softly	88
10.	Lips-Smile, laugh	94
11.	Lips- pointed	73
12.	Raised eyebrows	95
13.	Lowered eyebrows	73
14.	Parting eyebrows	87
15.	Joining eyebrows	69
16.	Scratching eyebrows	91
17.	Head shaking	89
18.	Head dropping	81
19.	Head nodding	102
20.	Hands on face or using them as supports	84
21.	Scratching on head	79
22.	Scratching on ears	68
23.	Scratching on nose	75

Table-4 shows the comparative performance of the Radial Basis Neural Network model [20], Hidden Markov (HMM)[38]model and Support Vector Machine (SVM)[37] model tested in this study. All these algorithms were implemented in Matlab and integrated with LabView vision module to process and classify image data. All the models were trained using facial feature's distance attributes from Cohn-Kanade database and also from a custom database which was populated using facial expressions of sample space of 30 students. The NN model [20] used in this study outperformed HMM and SVM for the Cohn-Kanade database in this study and proved to be most reliable given sample space is large.

Following facial expressions were targeted for training the proposed model.

a) Happy
b) Sad
c) Confused
d) Disturbed
e) Surprised

TABLE IV. CLASSIFICATION ACCURACY IN CUSTOM AND EXISTING DATABASES

No	Facial Expression	Classification Accuracy on Existing Database			Classification Accuracy on Custom Database		
		NN	*SVM*	*HMM*	*NN*	*SVM*	*HMM*
1	Happy	90.1 %	92.2 %	87.2%	81%	83%	74%
2	Sad	88.1 %	85.0 %	82.6%	78%	80.8%	78.2%
3	Confused	80.6 %	78.2 %	73.4%	77%	64.5%	72.2%
4	Disturbed	86.2 %	81.0 %	80%	72%	76.3%	69%
5	Surprised	85.3 %	84.7 %	84%	72%	71.4%	75%

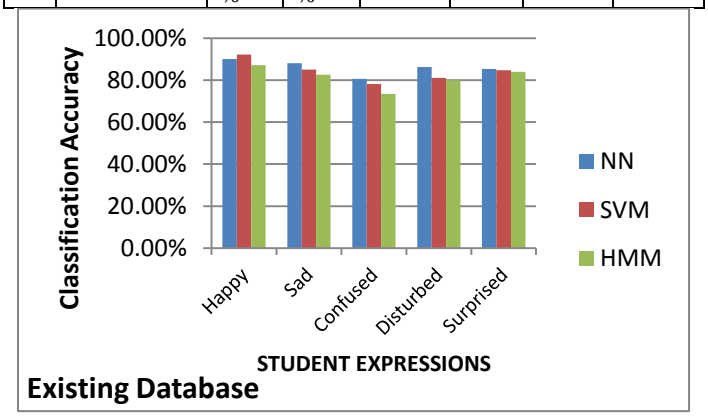


Fig. 3. Classification accuracy on existing database

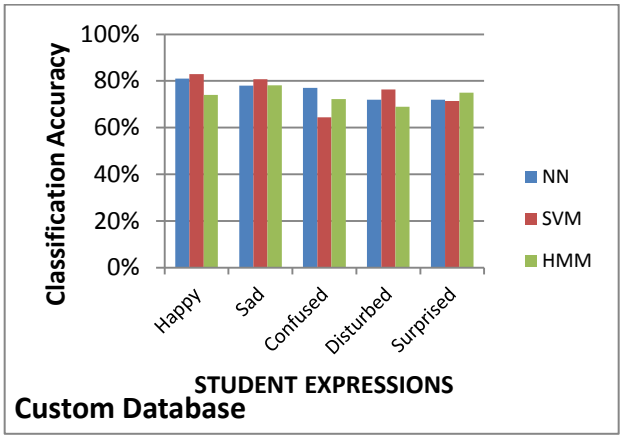


Fig. 4. Classification accuracy on custom database

Based on different distance sets, feature sets were populated and tested for reliability using Cohn-Kanade dataset. Expressions were readily classified with high accuracy, when test images were selected from the same database used for training our NN model. However, accuracy dropped but to acceptable level when custom image set of 30 students was used as a test data set. This problem can be circumvented by using a custom template for images and iteration of neural network training with on new datasets. A similar trend was observed for the SVM and HMM classification carried out on the Cohn-Kanade dataset. For the custom dataset, SVM and HMM results showed random classification rates which can be attributed to small sample space of facial features.

The response of the students was recorded by asking them to provide a score out of 10 for their attentiveness during each 10 min session and based on their feedback; attentiveness was correlated with the extracted facial expression sequences or simply mood patterns. Mood extraction was carried out during every 10 min session of the one hour mathematics lecture for all the 30 students. The total extracted mood patterns using Apriori and correct patterns based on the correlation results are shown in Table-5

TABLE V. VALID MOOD PATTERNS EXTRACTED DURING 6 X 10 MINUTE TIME FRAME IN ONE HOUR SESSION FOR 30 STUDENTS

Mood/Expression	Extracted Mood Patterns	Correct Mood Patterns	Percentage Error
Happy	122	92	75.4%
Sad	110	87	79.0%
Confused	125	89	71.2%
Disturbed	109	78	71.5%
Surprised	98	82	83.6%

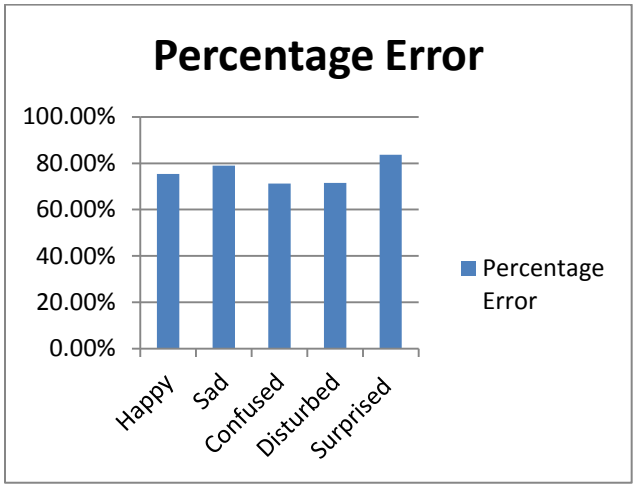


Fig. 5. Percentage error in Mood/ Facial expression classification

Results showed that mood patterns extracted had a high correlation with the feedback provided by the students. In all the cases, our proposed algorithm showed a success rate of over 70% in assessing the student's mood. This showed that a student's expression over a 10 minute timeframe can be used to predict and extract a student's mood, which in turn can be used to assess student's attentiveness in the class. The results showed that the proposed approach was very robust due to integration of neural network based classification and Apriori algorithm for mood extraction. The difference in success rates for each mood can be related to basic test settings, incomplete database and simpler NN training settings.

The proposed classification and mood extraction method do not attempt to address complete theory of emotions in context of e-learning, rather it is intended to devise a methodology in identifying any mood that is persistent is affecting a student's attention in an e-learning environment, which is an important consideration as highlighted in [31][32] for boredom, confusion and frustration states. The results provided in this research shows that the proposed technique is promising in assessing five moods in an active e-learning environment which were selected using a survey. The success percentage for assessing each emotional state is above 70%. In future work, more emotional states can be tested and based on the results from this study, a similar success ratio is expected given an extensive facial feature database is used.

V. CONCLUSION

The art of understanding how different students comprehend educational content during an online study session requires detailed investigation on the behavior and emotional

state of the student throughout the lecture [33][34]. This research was carried out to determine possible ways to observe and analyze behavior of a student with an aim to understand the events triggering his emotional detachment during an online class.

Visual data acquired using high definition cameras contains a lot of information when stored over a long period of time, and it needs to be continuously recorded thus accumulating into very large data. Data mining approaches can help in similar can help in mining patterns from such large data. Important rules based on correlation characteristics of classification attributes like distance can be acquired to characterize mood swings and changes that affect the learning curves of a student in an e-learning environment. The results can help in better understanding the complete eco-system of an E-learning environment where learner-machine active interaction and raised level of student's engagement is of prime concern. The delivery of e-learning content as well as attitude discrepancies in a student can be then adequately addressed to enhance the student's involvement and attentiveness during e-leaning. This can be done during or after the e-learning session based on the preference of the student and/ or the E-learning administrator

Main contribution of this research is the integrated approach with neural network facial recognition and Apriori based mood extraction, which showed a probability of over 70% for detecting 5 emotional states or moods.

Facial expressions describe the emotional state of the learner and analysis of content and delivery methods can be carried out to achieve optimal experience in an e-learning environment [35]. However it is difficult to devise universal standard content delivery systems for every learner, therefore, specific testing sessions may be incorporated in an e-learning system to allow customization as per student's learning curves.

The results assimilated using a survey response from various academics showed that facial features are the best method to observe changes in the mood of a student and relevant causes can be extracted by relating the timeline with the content delivery and student's successive changes in facial expressions. The problem addressed in this study is limited to determining how a mood can be extracted by associating feature sets comprised of various facial expressions. The cause of alterations in mood and mental state are altogether another problem and is not discussed in this research.

Mental state of a student can be observed using his facial expressions as facial features tend to change and provide the best depiction of what student have in mind [36]. As involvement of a student during unsupervised learning is critical in improving his learning potential, it is pertinent to learn the problems faced by students in an e-learning environment.

Finally, most contribution of this research lies in the results which showed that facial expressions extracted using neural networks, and the reduction of sample space using Apriori algorithm can be actively used to derive student's emotional state during content delivery in an e-learning system. The proposed integrated approach showed a high probability of positive mood detection rate (>70%) for five moods. Happy, sad, confused, disturbed, and surprised moods or emotional states

For future work, the resultant data can be used to optimize e-learning content delivery to engage learner more actively in real-time when a mood leading to inattentiveness is detected.

REFERENCES

[1] G. Eason, B. Noble, and I. N. Sneddon, "On certain integrals of Lipschitz-Hankel type involving products of Bessel functions," Phil. Trans. Roy. Soc. London, vol. A247, pp. 529–551, April 1955. *(references)*\

[2] M. Nur-Awaleh and L. Kyei-Blankson, "Assessing E-learning and student satisfaction in a blended and flexible environment," 2010 International Conference on Information Society, London, 2010, pp. 481-483.

[3] C. Leghris and R. Mrabet, "Cost Comparison of E-Learning Solutions," 2006 7th International Conference on Information Technology Based Higher Education and Training, Sydney, NSW, 2006, pp. 817-824.

[4] Fresen, J. (2007). A taxonomy of factors to promote quality web-supported learning. International Journal on E-Learning, 6(3), 351-362.

[5] J. Yu, "An Infrastructure for Real-Time Interactive Distance E-Learning Environment," 2009 First International Conference on Information Science and Engineering, Nanjing, 2009, pp. 3219-3222.

[6] Mohamed Sathik M, Sofia G (2011) Identification of student comprehension using forehead wrinkles. 2011,International Conference on Computer, Communication and Electrical Technology (ICCCET), pp 66–70.

[7] N. Fragopanagos and J. G. Taylor, ―2005 Special Issue: Emotion recognition in human-computer interaction. Neural Networks, Neural May 2005. Networks-Special Issue: Emotion and Brain, vol. 18, pp. 389–405.

[8] Rothkrantz, L.J.M. E-learning in virtual environments, Communication & Cognition, Vol 42, No. 1&2, pp 37-52, 2009.

[9] A. Walia, N. Singhal and A. K. Sharma, "A Novel E-learning Approach to Add More Cognition to Semantic Web," 2015 IEEE International Conference on Computational Intelligence & Communication Technology, Ghaziabad, 2015, pp. 13-17.

[10] Fabri, M., Moore, D.J., Hobbs, D.J (2004) "Mediating the Expression of Emotion in Educational Collaborative Virtual Environments: An Experimental Study", in International Journal of Virtual Reality, Springer Verlag, London

[11] M. Feidakis, T. Daradoumis, S. Caballé and J. Conesa, "Measuring the Impact of Emotion Awareness on e-learning Situations," 2013 Seventh International Conference on Complex, Intelligent, and Software Intensive Systems, Taichung, 2013, pp. 391-396.

[12] R. Nkambou, (2006) "Towards Affective Intelligent Tutoring System", Workshop on Motivational and Affective Issues in ITS. 8th International Conference on ITS 2006, pp 5-12

[13] Chaffar, S. and Frasson, C. (2005). "The Emotional Conditions of Learning". Proceedings of the FLAIRS Conference 2005, .pp. 201-206

[14] Guey-Shya Chen and Min-Feng Lee, "Detecting emotion model in e-learning system," 2012 International Conference on Machine Learning and Cybernetics, Xian, 2012, pp. 1686-1691.

[15] Bailenson J, Beall A, Blascovich J, Raimundo M, Weishbush M (2000) "Intelligent agents who wear your face: User's reactions to the virtual self" Technical Report, Center for the Virtual Environment and Behaviors Department of Psychology, University of California, Santa Barbara

[16] R. Brunelli and T. Poggio, ' Faical Recognition: Features versus Templates," IEEE Trans. Pattern Analysis and Machine intelligence, vol. 15, no.10, pp. 1042-1052, Oct. 1993.

[17] LA. Essa and A.P. Pentland, "Coding, Analysis, Interpretation, and recognition of facial expressions,"IEEE Trans. Pattern Analysis and Machine Intelligence, vol. 19, no.7, pp-757-763, July 1997

[18] J.W.Grzymala-Busse, "On the unknown attribute values in learning from examples," in Proceedings of the ISMIS-91, 6th International

Symposium on Methodologies for Intelligent Systems, Lecture Notes in Artificial Intelligentce, Vol.542,Springer-Verlag, Berlin Herdelberg New York,1991,pp.368-377.

[19] G. Zhang, "Neural networks for classification: a survey", IEEE Trans. Syst., Man, Cybern., Syst., vol. 30, pp. 1094–6977, Nov 2000.

[20] Consortium.ri.cmu.edu, 'Cohn-Kanade (CK and CK+) database Download Site', 2015. [Online]. Available: http://www.consortium.ri.cmu.edu/ckagree/. [Accessed: 20- Sep- 2014]

[21] Weihua Wang, "Face Recognition Based on Radial Basis Function Neural Networks," Future Information Technology and Management Engineering, 2008. FITME '08. International Seminar on , vol., no., pp.41,44, 20-20 Nov. 2008

[22] J. Han, M. Kamber,"Data Mining: Concepts and Techniques", Morgan Kaufmann Publishers, Book, 2000.

[23] F. H. AL-Zawaidah, Y. H. Jbara, and A. L. Marwan, "An Improved Algorithm for Mining Association Rules in Large Databases," Vol. 1, No. 7, 311-316, 2011

[24] T. C. Corporation, "Introduction to Data Mining and Knowledge Discovery", Two Crows Corporation, Book, 1999

[25] R. Srikant, "Fast algorithms for mining association rules and sequential patterns," UNIVERSITY OF WISCONSIN, 1996.

[26] S. Rao, R. Gupta, "Implementing Improved Algorithm Over APRIORI Data Mining Association Rule Algorithm", International Journal of Computer Science And Technology, pp. 489-493, Mar. 2012

[27] Ekman, P. & Oster, H. (1979). Facial Expressions of Emotion. Annual Review of Psychology, 30, 527-554.

[28] Ekman, P., Friesen, W. V., & Ancoli, S. (1980). Facial Signs of Emotional Experience. Journal of Personality and Social Psychology, 39(6), 1125-1134.

[29] Ekman, P. (1993). Facial Expression and Emotion. American Psychologist, 48(4), 384-392.

[30] Ekman, P. & Keltner, D. (1997). Universal facial expressions of emotion: An old controversy and new findings. In Segerstråle, U. C. & Molnár, P. (Eds.), Nonverbal communication: Where nature meets culture (pp. 27-46). Mahwah, NJ: Lawrence Erlbaum Associates.

[31] Aghababyan, Ani. "E3: Emotions, Engagement and Educational Games." InEducational Data Mining 2014. 2014.

[32] Craig, S. D., Graesser, A., Sullins, J., & Gholson, B. (2004). Affect and learning: An exploratory look into the role of affect in learning. Journal of Educational Media (now: Learning, Media & Technology), 29, 241–250.

[33] D'Mello, S. K., Craig, S. D., & Graesser, A. C. (2009). Multimethod assessment of affective experience and expression during deep learning. International Journal of Learning Technology, 4, 165–187.

[34] T. V. Pham, M. Worring, and A. W. M. Smeulders. Face detection by aggregated bayesian network classifiers. Pattern Recogn. Lett., 23(4):451–461, 2002

[35] Li Xia, "Facial Expression Recognition Based on SVM," in Intelligent Computation Technology and Automation (ICICTA), 2014 7th International Conference on , vol., no., pp.256-259, 25-26 Oct. 2014

[36] S. Deshmukh, M. Patwardhan and A. Mahajan, "Survey on real-time facial expression recognition techniques," in IET Biometrics, vol. 5, no. 3, pp. 155-163, 9 2016.

[37] A. W. P. Fok, H. S. Wong and Y. S. Chen, "Hidden Markov Model Based Characterization of Content Access Patterns in an e-Learning Environment," 2005 IEEE International Conference on Multimedia and Expo, Amsterdam, 2005, pp. 201-204.

[38] W. Gong and W. Wang, "Application research of support vector machine in E-Learning for personality," 2011 IEEE International Conference on Cloud Computing and Intelligence Systems, Beijing, 2011, pp. 638-642.

Impact of Domain Modeling Techniques on the Quality of Domain Model: An Experiment

Hiqmat Nisa
Dept. of Computer Science & Software Engineering
International Islamic University
Islamabad, Pakistan

Muhammad Uzair Khan
Dept. of Computer Science National University of Computer & Emerging Sciences (FAST-NU)
Islamabad, Pakistan

Salma Imtiaz
Dept. of Computer Science & Software Engineering
International Islamic University
Islamabad, Pakistan

Saima Imtiaz
Dept. of Computer Science & Software Engineering
International Islamic University
Islamabad, Pakistan

***Abstract*—The unified modeling language (UML) is widely used to analyze and design different software development artifacts in an object oriented development. Domain model is a significant artifact that models the problem domain and visually represents real world objects and relationships among them. It facilitates the comprehension process by identifying the vocabulary and key concepts of the business world. Category list technique identifies concepts and associations with the help of pre defined categories, which are important to business information systems. Whereas noun phrasing technique performs grammatical analysis of use case description to recognize concepts and associations. Both of these techniques are used for the construction of domain model, however, no empirical evidence exists that evaluates the quality of the resultant domain model constructed via these two basic techniques. A controlled experiment was performed to investigate the impact of category list and noun phrasing technique on quality of the domain model. The constructed domain model is evaluated for completeness, correctness and effort required for its design. The obtained results show that category list technique is better than noun phrasing technique for the identification of concepts as it avoids generating unnecessary elements i.e. extra concepts, associations and attributes in the domain model. The noun phrasing technique produces a comprehensive domain model and requires less effort as compared to category list. There is no statistically significant difference between both techniques in case of correctness.**

***Keywords*—Domain Model; UML; Experiment; Noun Phrasing Technique; Category List Technique**

I. INTRODUCTION

UML (Unified modeling language) is gaining fame since its inception in 1997; it is being commonly practiced by the industry to model object oriented software systems. UML plays a significant role in reducing the complexity of large software system by modeling different aspects throughout SDLC phases. Object Oriented Analysis (OOA) is carried to understand and model the problem domain in the form of real world objects, which can later be translated into the solution. It describes problem domain from the perspective of objects and emphasizes on identifying and describing the concepts, attributes and associations in the problem domain [1]. One of the main outcomes of OOA is a domain model which models the problem domain objects along with their associations and attributes.

Domain model is one of the most important UML artifact used to understand the problem domain. It represents vocabulary and key concepts, important to the business world [1] [2] [3]and consists of visual representation of concepts, attributes and association among conceptual classes in the real world domain. It also presents general vocabulary, which helps in clear communication between the team members and helps elevate the level of understanding between the development team and customer side [2] [3]. A solution which is representative of the customer needs requires a domain model that is representative of the domain. A clear and precise domain model can also help in reducing risk [4] and effort and cost of rework required at later stages [5]. Therefore one of the major goals of OOA is to create an accurate and complete domain model.

The domain model can be created using two different techniques suggested by Larman [1]: category list technique and noun phrasing technique. To identify potential candidate classes and associations; category list technique provides a list of categories which are usually important to business information systems. Each category represents entities or concepts related to real-world. Sets of candidate classes produced by all categories are quite independent from each other whereas, noun phrasing technique is linguistic analysis. Noun phrasing technique involves the identification of nouns and noun phrases in the domain description, and considers them as conceptual classes or attributes [1]. These techniques have not been empirically evaluated for their effectiveness in creating a quality domain model. Therefore an experiment was performed to evaluate the effectiveness of techniques in creating a complete and accurate domain model. The experiment was conducted with help of undergraduate students of fourth semester of software engineering, as they are assumed to be familiar with the models and notations of UML. This experiment is focused to answer the below given research questions.

RQ1: What is the effect of noun phrasing and category list technique on the quality of the domain model?

RQ2: What is the amount of effort required to create the domain model using both techniques?

The quality of the domain model is determined on the basis of completeness and correctness of the domain model, whereas the amount of effort is measured in terms time taken to create the model. The rest of the paper is organized as follows: Section 2 presents the Background and Related Work. In Section 3 elaborates on the Design of the Experiment and Section 4 discusses the Analysis and Results. Finally conclusion and future work is given in section 5.

II. Background and Related Work

Domain model is the most important and common model in object oriented analysis. It describes the noteworthy concepts or objects in problem domain. It is a representation of the real-world conceptual classes, attributes of the classes and associations among them Domain model is an improved version of the project dictionary, where the terms used in the project are present along with the graphical visualization of the connections between them. It can be termed as a simplified version of a class diagram, one that does not incorporate responsibility assignment [1]. Most of the conceptual classes modeled in domain model become part of the class diagram, which are important to software development [2] [6].

Domain model can be created using two different techniques namely: noun phrasing technique and category list technique [1]. There are some basic steps involved to create a domain model i.e. identification of conceptual classes along with their attributes and associations and unnecessary candidate classes.

Noun phrasing technique uses grammatical analysis of use case description to identify nouns and noun phrases and consider them as candidate conceptual classes or attributes. For the identification of associations, verb phrases are identified between entities and are considered as relationships between conceptual classes. However, for the identification of potential candidate classes and associations using category list technique Larman [1] provides a list of categories which are usually important to business information system and also provides guideline to eliminate useless concepts which are not appropriate to be implemented

Noun phrasing technique is the simplest approach to create domain model, but result in many imprecision problems e.g. words may be ambiguous or the identification of redundant classes due to synonyms in use case description and noun phrase may also be an attributes rather than a concept [1].Identifying noun and noun phrases is an analyst's job to examine each noun phrase and consider it either as a concept or an attribute. Some guidelines have been proposed by Larman to identify and refine attributes. The research focuses on empirically evaluating both of the techniques to observe their effect on the quality of domain model.

The Literature survey highlights that various empirical studies have been conducted to evaluate the impact of different techniques used to construct different UML models. Most of the target UML models are use case diagram, Class diagram and sequence diagram. The work of T. Yue et.al. [7] for instance, investigated whether restricted use case modeling (RUCM) approach or traditional use case template produced high quality analysis models i.e. Class diagram and sequence model. Subjects designed a class and sequence diagram of a given software systems using RUCM approach and traditional use case template. Results pointed out that RUCM produced better quality model than traditional use case template. Similar experiment was performed S.Tiwari *et al.* [8] [9], where they investigated the impact of use case templates on the quality of class diagram and use case diagram. They concluded [9]that no template is statistically significant better over another in terms of completeness, consistency, understandability, redundancy and fault proneness. However formal use case template produced high quality class diagram as compared to UML use case template and formal use case produced less redundant elements in class diagram [10]. Another study l [11], evaluated the effectiveness of two techniques i.e. validation and derivation technique on the quality of class diagram, and concluded that derivation technique produced more complete class diagram as compared to validation technique.

The quality of domain model is also evaluated by some researchers. The impact of system sequence diagram (SSD) and system operation contract (SOC) is observed on the quality of domain model [12]. The subjects designed domain model with SSD and SOC and without SSD and SOC. Two factors were involved to evaluate the quality of domain model, i.e., completeness and time. Author concluded that using SSD and SOC to construct a domain model, improves the quality of domain model in case when subjects have enough practice to take advantage from SSD and SOC. Another study conducted by S. Espana et al. [10]evaluated the quality of conceptual model constructed by two alternative techniques i.e. text-based derivation technique and communication based derivation technique. Participants were required to construct conceptual model using two alternative techniques. The quality of derived conceptual model was evaluated according to completeness and number of faults present (model validity) in participants conceptual model. The results highlight that the participants who used communication based derivation techniques produced 9.22% more complete conceptual diagram as compared to those who used text based derivation technique. Briand et al. [12] investigated that whether the use of SSD or SOC in domain model construction, improve the quality of domain model or not. Whereas, a main concern is to evaluate domain model construction technique suggested by Larman [1].

Most of the researchers conducted empirically studies, to compare different techniques [10] [7] [11] [13]for the purpose that which technique leads to high quality UML diagram. However most of the target UML models are class, use case diagram and sequence diagram. This research is focused on

the quality of resultant domain model created via noun phrasing and category list technique.

III. Experiment Planning

The research is validated with help of an experiment .This section explains design of the experiment. The experimental guidelines were followed to design the experiment in a controlled environment as suggested by C. Wohlin [14]. All the steps of an experiment to evaluate the quality of domain modeling techniques are reported in this section.

A. Experiment Definition

The purpose of this research is to empirically evaluate the impact of noun phrasing and category list technique on the quality of domain model. Our main concern is the creation of a domain model by the subjects via noun phrase or category list technique. As a result, two treatments are described as independent variable. One describes the creating of domain model using noun phrasing technique, and the other one describe the domain model using category list technique. The aim of this experiment is to evaluate the quality of domain model in terms of correctness, completeness, and effort required to design a complete domain model.

B. Context selection and subject

The selection of the subjects is very important for generalizing the results of experiment. Results generalization can be achieved by satisfactory sample size and random subject selection [14]. This experiment is conducted with 68 fourth year undergraduate computer science students in a famous Science and technology University of Islamabad, Pakistan. The students are familiar with UML notation and domain modeling techniques. They studied UML as part of their software engineering course in initial semesters. All the students have similar experience in modeling UML diagrams. The students were selected as experiment subjects as they fulfill the criteria i.e. participants who have similar education background, adequate knowledge and training of domain modeling.

To avoid biasness simple random sampling [14] is used for subject selection, i.e. subjects are selected from the population at random. Subjects were divided into two groups: group A and group B according to their grades. The categorization of students in two groups according to their grades is done to minimize the impact of students' capability on experiment's results. Before conducting the experiment a brief presentation is given to students about domain modeling techniques and the experiment. However the hypothesis of the experiment is not disclosed.

Two different systems were used as objects in this experiment, Automatic Teller Machine (ATM) and internet book store system (IBS). The ATM use case describes the process of withdraw fund and card verification as discussed in [15]. The IBS system purchases books over internet via credit card and Amazon website as discussed in [16]. We provide the experimental systems of limited complexity due to time constraints, so that subjects are able to finish their task.

C. Dependent and independent variable

There are two independent variables, Technique (category list and noun phrase) and Domain used (ATM and IBS).

Quality of domain model is evaluated by three dependent variables i.e. completeness, correctness and effort. Correctness is calculated in terms of average value of Useless Concepts(UC), Missing Concepts (MC), Extra Relationships (ER), Missing Relationships (MR), Extra Attributes(EA), Missing Attributes (MA) and Missing Generalizations(MG [12]. Completeness is defined as average of correctly identified elements in the domain model i.e. average number of Correct Concepts (CC), Correct Relationships (CR) and correct attributes (CA) and Correct Generalizations (CG) [7]. Table I and table II present the completeness of domain model completeness.

The second dependent variable checks the significant difference between the effort required to design a domain model by subjects who use noun phrase technique and those who use category list technique. The effort is calculated in terms of time, measured in minutes. Only that time was considered which utilized in creation of fully completed or partially completed domain model. The time is computed by subtracting the start time of the experimental task from end time of the experimental task.

D. Hypothesis

Two main research questions are investigated in this experiment. The first question contains a number of hypotheses shown in table III. According to experimental design one independent variable was considered called method, with two treatments: category list technique and noun phrasing technique, and three dependent variables correctness, completeness of domain model and effort required to complete a domain model. Thus two tailed hypothesis i.e. alternate and null hypothesis was formulated. The null hypothesis (H0) for each dependent variable is: there is no difference between category list technique and noun phrasing technique in terms of completeness and correctness of domain model and required effort. The alternative hypothesis (H1) is defined as: category list technique produces different quality of domain model, or different effort is required to complete a domain model when compared to noun phrasing technique.

E. Experiment Design

Crossover design is followed in the experiment. Crossover design is a repeated measurement design such that each subject receives different treatments during different time periods. This experiment is conducted in two labs. In first lab, subjects in group A are required to design a domain model for ATM system using noun phrasing technique and group B have to construct a domain model for Internet book store system using category list technique. In second lab, same subjects of group A are required to complete the domain model for Internet book store system using noun phrasing technique and same subjects of group B are required to design a domain model using category list technique for ATM System depicted in table IV. A short presentation was given to the participants to introduce the domain model and its concepts along with the procedure of the

experiment. The hypothesis of the research was not disclosed to avoid any biases later on. The participants were given 40-45 minutes to finish the domain model.

The experiment is performed in supervision of the lab supervisor in both labs. All the required material is provided to the participants. The participants were required to note the time before starting the experiment and after completion of the experiment. Participants were required to construct the domain model using one technique in first half and alternative technique in second half, respective data is collected.

TABLE I. MEASURES USED TO DERIV DOMAIN MODEL

No#	Measures	Specification
	NCref	Number of correct Concepts in Reference model
	NRref	Number of correct Relationships in Reference model
	NAref	Number of correct Attributes in Reference model
	NGref	Number of correct Generalizations in Reference model
	NCC	Number of correct Concepts in Subjects model
	NCR	Number of correct Relationships in Subjects model
	NCA	Number of correct Attributes in Subjects model
	NCG	Number of correct Generalizations in Subjects model
	NUC	Number of useless Concepts in Subjects model
	NER	Number of extra Relationships in Subjects model
	NEA	Number of extra Attributes in Subjects model

TABLE II. QUALITY MEASURES FOR DOMAIN MODEL

Dependent		Formula
Completeness	Class Completeness *Ccom= NCC/ Ncref*	Completeness= *(Ccom+ Rcom+ Acom+ Gcom)/4*
	Relationships Completeness *Rcom= NCR/NRref*	
	Attributes Completeness *Acom= NCA/NAref*	
	Generalizations Completeness *Gcom= NCG/NGref*	
Correctness	Number of Useless Concepts *NUC*	Correctness= *(NUC+NMC+NER+NMR+NEA+ NMA+ NMG)/7*
	Number of Missing Concepts *NMC= Ncref – NCC*	
	Number of Extra Relationship *NER*	
	Number of Missing Relationship *NMR= NRref – NCR*	
	Number of Extra Attributes *NEA*	
	Number of Missing Attributes *NMA= NAref –NCA*	
	Number of Missing Generalizations NMG	

TABLE III. HYPOTHESIS FOR DOMAIN MODEL CORRECTNESS, COMPLETENESS AND REQUIRE EFFORT

Dependent variable	Null Hypothesis	Alternative Hypothesis
Correct Concepts (CC)	CC(CLT)=CC(NPT)	CC(CLT) ≠CC(NPT)
Use Concepts (UC)	UC(CLT)=UC(NPT)	UC(CLT) ≠UC(NPT)
Missing Concepts (MC)	MC(CLT)=MC(NPT)	MC(CLT) ≠MC(NPT)
Correct Relationships(CR)	CR(CLT)=CR(NPT)	CR(CLT) ≠CR(NPT)
Extra Relationships(ER)	ER(CLT)=ER(NPT)	ER(CLT) ≠ER(NPT)
Missing Relationships(MR)	MR(CLT)=MR(NPT)	MR(CLT) ≠MR(NPT)
Correct Attributes (CA)	CA(CLT)=CA(NPT)	CA(CLT) ≠CA(NPT)
Extra Attributes (EA)	EA(CLT)=E(NPT)	EA(CLT) ≠E(NPT)
Missing Attributes(MA)	MA(CLT)=MA(NPT)	MA(CLT) ≠MA(NPT)
Correct Generalizations(CG)	CG(CLT)=CG(NPT)	CG(CLT) ≠CG(NPT)
Missing Generalization(MG)	MG(CLT)=MG(NPT)	MG(CLT) ≠MG(NPT)
Overall Completeness (Com)	Com (CLT)= Com (NPT)	Com (CLT) ≠Com (NPT)
Overall Correctness (Corr)	Corr (CLT)= Corr (NPT)	Corr (CLT) ≠ Corr (NPT)

Effort	Time(CLT)=Time(NPT)	Time(CLT) ≠Time(NPT)

1) Co factors: There are some extraneous factors that affect the experiment results. These are also known as co founding variables that can also affect the results. In case of influence it becomes difficult to infer that the results are due to the independent variable or due to these co-founding variables. These extraneous factors must be minimized to increase the experiment's effectiveness. In this research students' ability and system complexity are considered as co-founding variables. Subjects of the experiment were choose from the same batch i.e. 4th year students to ensure same level of knowledge and skills regarding domain modeling, however we cannot ignore the fact that students belonging to same class may have different analytical and design skills. These skills would also affect the design of domain model from different complexity systems. Therefore we used a block design of experiment to control the impact of these co-founding variables on the output of the experiment.

Subjects were divided into two blocks according to their grades in software engineering course, so that each group consists of students with almost the same ability as far as software engineering knowledge and skills is concerned.

2) Learning and fatigue effect: When subjects deal with the same problem more than once, their response will be better at the second exposure as compared to first one, because human learn from previous experience. As a result any significant changes in the second time can be the effect of practice or learning [14].

In experiment, subjects were required to complete a domain model twice. Different system was used in second half to avoid learning effect.

TABLE IV. EXPERIMENT DESIGN

Lab	**Task**	**Group A**	**Group B**
Lab1	Domain ATM	Category list	Noun phrase
Lab 2	Domain IBS	Noun phrase	Category list

F. Instrumentation

There are three types of instruments associated with experiment: experimental objects, guidelines and measurement [14].

Experimental objects can be a document or source code on which subjects have to work. During experiment planning it is necessary to select appropriate objects i.e. in this experiment; use case description is required for the creation of domain model. In this experiment objects consist of use case description of both software systems (ATM, IBS). Use case description of ATM system [15] and IBS system [16] were selected from literature. A document was provided to students which contains a brief use case description and students were required to design a domain model using pen and papers.

Regarding experiment guidelines, a brief presentation is given to the students in the beginning of the experiment. In which the students were briefly explained about the list of documents provided, the task to be performed, and the submission strategy. A written instructions document is also given which students return at the end of the experiment. The students were allowed to ask questions before start of the experiment. The students were required to complete the domain model within 45 to 50 minutes. This time selection to construct a domain model is based on the pilot study performed during course work activity.

Measurements contain, documents prepared to collect data and evaluation criteria to compute dependent variables. The use case description documents were prepared and validated. We compared students' domain model with reference model to measure the correctness and completeness of students' domain model. The reference domain model is design by external party, which consists of three researchers having 5 to 10 years of experience in UML and software engineering. The following criteria are followed to evaluate the students' domain model.

- All the concepts were considered correct if different names were used by students for the specific concept in reference model.
- All the relationships belonging to Missing concept in the reference model were considered missing.
- All the relationships of extra concepts were not considered as extra relationships.
- Attribute identified for extra concepts were not considered as extra attributes.
- Attributes were considered as extra identified attributes which are defined in the wrong concept.
- We assume the missing multiplicity to be one.
- In the inheritance, if super class is missing in the students' model, and attributes of super class is correctly defined in the sub class, then those attributes were considered as correctly identified attributes. And missing super class relationships were also considered correct if sub class is correctly associated with the class having direct relationship with super.

G. Analysis Procedure

Data analysis procedure consists of three dependent variables (domain model Correctness, completeness and effort involved to design a domain model), and one independent variable (Method), with two treatments (noun phrase technique and category list technique). The data analysis is performed with help of statistical test. Descriptive statistics presents the initial picture of collected data. Descriptive statistics summarize and presents the quantitative description in an effective way. Some basic descriptive statistics like, mean, standard deviation, minimum and maximum values were presented.

A Mann Whitney U-test was performed for each task related to designing a domain model to compare the means of dependent variables. The dependent variables are not normally distributed therefore we have selected Mann Whitney test which overcame the data normalization assumption.

Three-way ANOVA test is used to analyze combined data collected from lab 1 and lab 2 and extraneous factors which influence the dependent variables. It is used to identify the significance of main effect i.e.: the effect of and interaction between factors [17]. In this experiment two extraneous factors are considered, software systems and students' ability. The purpose of considering these two factors is to analyze the effect of systems' complexity and students' level of understanding on dependent variables and identifying possible interaction between factors.

H. Validity threads

1) Internal validity

Internal validity is concerned with cause-effect relationship among different variables .Internal validity threats can be present when the results of experiment are influenced by extraneous factors like learning and fatigue effect. Learning and fatigue effect is mitigated using cross-over experiment design and two different systems used in different labs.

Although students have same background knowledge but based on their ability subjects were divided into two balance groups according to their grades.

2) Construct validity

Construct validity threats are concerned with the relationship between concepts and construct being studied (correctness and effort). The measurement criteria were briefly explained. We believe that these measurements are reliable. The time factor is directly related to effort being used. Correctness and completeness cover all the domain model elements.

3) External validity

There are two major external validity threats which are related to this experiment, and these threats are usually associated with controlled experiment because of artificial environment used. They are: Are the sample of subjects in this experiment representative of software professionals? Is the material used in experiment representative of real software industry system in terms of complexity and size?

Regarding issue one, 4th year undergraduate student have acceptable knowledge about software engineering and UML modeling. They also practice UML and software engineering concepts during their assignments and projects. Their experience is almost same as junior professionals. Secondly, our purpose is to find the effectiveness of domain modeling techniques which do not need such a high level programming skills and experience. Students do not have exposure about different domain as professionals, but they are familiar about the domain modeling techniques and their usage, which they can apply on any problem domain.

Regarding the second issue, Software systems used in this experiment are small as compared to industrial software systems, because it is not feasible to take large industrial system in limited time [18], but its size and complexity is comparable with other systems used in related experiments [7] [9] [12].

4) Conclusion validity

Conclusions validity threats are related with issues that influence the capability to draw a correct conclusion about experimental hypothesis based on experimental results. Regarding this experiment, appropriate statistical tests were performed to find statistically significant difference. In case where little difference is found but not significant, power analysis was performed to avoid accepting false null hypothesis.

IV. Analysis

Table V and table VI show a Mann-Whitney test results. Overall results show a lack of significant difference between two groups in different dependent variables.

In lab 1, we see a significant difference in the correct concepts (p-value= .021) and correct generalizations (p-value= .039) dependent variables only. From the mean rank of correct concepts show that students produced more correct concepts using category list technique than noun phrase technique.

In lab 2, a significant difference is shown in correct attributes (p-value=.000) and overall completeness (p-value=.009) only. No other dependent variables show significant differences. According to mean rank, subjects produced more correct attributes using noun phrase as compared to category list technique.

Table VI shows the results of overall correctness and effort. As discussed in the dependent variables section that the overall correctness is calculated as average of all the extra dependent variables (concepts, relationships and attributes) and all the missing dependent variables (concepts, relationships and attributes). Lower the overall correctness mean, better will be the quality of domain model.

We can observe from the table VI that those students who used category list technique produced more missing concepts as compare to those who used noun phrase technique in lab 1. However, in lab 2 a significant difference is found in extra attributes (p-value=.000) and missing attributes (p-value=.0000). It can observe from the value of mean rank, that noun phrase technique produced more numbers of correct elements of domain model (concepts, relationships and attributes), However it also produced large number of extra elements in the domain model. No significant difference is found in overall Correctness dependent variable.

We also conduct a power analysis to determine the power of those statistical tests having no significant results. Before accepting null hypothesis we compare minimum effect size required to obtain 80% power with observe effect size. In case of ATM software system, the minimum effect size required to obtain 80% power for overall completeness is 0.512 but the observed effect size is 0.432. Due to the small effect size the observed power is 70%. So we cannot provide any erroneous conclusion about overall completeness of domain model. On the other hand, the observed power of overall correctness is also very small for IBS system. So we cannot reject the null hypothesis.

According to second research question, a significant difference is observed between effort require in term of time. In lab 1 (p-value=.000) and in lab 2 (p-value=.004) were observed for required effort. So from the mean rank we can say that students spent more time in designing a domain model using category list technique as compare to noun phrase technique.

We apply three-way ANOVA test to analyze the combine data of lab 1 and 2 and possible interaction of co-factors, shown in table VIII. In this experiment, System and Ability factors are considered. We observe a significant main effect for the System factor in overall completeness and overall correctness. This significant main effect is in favor of ATM system. The reason of main effect of system may be that students feel more comfortable and performed better in ATM system. We also observe that noun phrase technique produced 6% more complete domain model as compared to category list technique. We do not found any significant interactions between System and Method, Ability and Method, System and Ability. Which is further elaborated on interaction plot.

Interaction plots highlight interaction in case of nonparallel lines, whereas parallel lines indicate no interaction at all. It can be seen from figure (a) and (b) that subjects with high and low ability performed similarly in both systems in case of completeness of domain model. However it can also be seen that subjects with high ability were able to make a more complete domain model in ATM system using noun phrasing technique. Regarding correctness it is observed from figure (c) and (d) that high and low ability students performed the same whether they used noun phrasing technique or category list in both software system.

Regarding required effort, we observed a significant time difference to complete the domain model in case of both systems. In both software systems students spent more time to complete a domain model using category list technique as compared to noun phrasing technique. This is also observed from interaction plot (e) and (f).

TABLE V. MANN-WHITNEY U TEST OF OVERALL COMPLETENESS

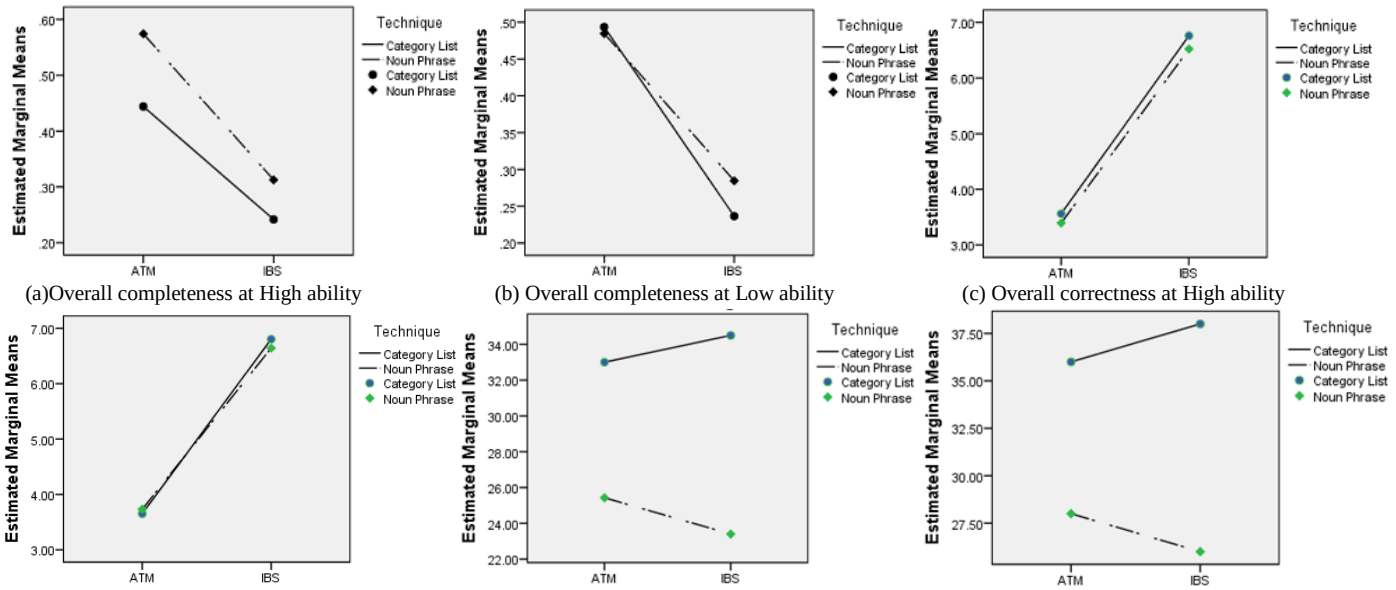
Dependent variable	**Technique**	**Lab 1**			**Lab 2**		
		Mean	***Mean Rank***	***P-value***	***Mean***	***Mean Rank***	***P-value***
Correct Concepts	Noun phrase	5.34	28.4	**.021**	7.08	32.3	.355
	Category List	5.81	38.5		7.41	36.6	
Correct Relationships	Noun phrase	3.42	34.8	.561	5.79	35.5	.664
	Category List	3.12	32.1		5.23	33.4	
Correct Attributes	Noun phrase	3.45	36.1	.251	3.28	45.8	**.000**
	Category List	2.66	30.8		.882	23.1	
Correct Generalization	Noun phrase	2.36	37.5	**.039**	1.17	36.1	.445
	Category List	1.72	29.4		.794	32.8	
Overall Completeness	Noun phrase	0.53	36.9	.144	0.29	40.7	**.009**
	Category List	0.46	30.05		0.23	28.2	

TABLE VI. MANN-WHITNEY U TEST OF OVERALL CORRECTNESS AND EFFORT

Dependent Variable	**Technique**	**Lab 1**			**Lab 2**		
		Mean	***Mean Rank***	***P-value***	***Mean***	***Mean Rank***	***P-value***
Useless Concepts	Noun phrase	1.15	36.8	.132	2.50	34.8	.880
	Category List	.88	30.1		2.38	34.1	
Missing Concepts	Noun phrase	2.65	28.4	**.021**	7.91	36.6	.355
	Category List	2.18	38.5		7.58	32.3	
Extra Relationships	Noun phrase	2.33	36.1	.254	3.02	34.8	.876
	Category List	2.00	30.8		3.05	34.1	
Missing Relationships	Noun phrase	5.57	32.1	.561	13.20	33.4	.664
	Category List	5.87	34.8		13.76	35.5	
Extra Attributes	Noun phrase	4.03	34.9	.404	2.00	45.0	**.000**
	Category List	3.69	31.0		.352	23.9	
Missing Attributes	Noun phrase	8.54	30.7	.323	12.41	23.1	**.000**
	Category List	9.33	35.3		15.11	45.8	
Overall Correctness	Noun phrase	4.66	32.97	.822	6.57	30.63	.105
	Category List	3.55	34.03		6.78	38.73	
Effort	Noun phrase	3.60	8.40	**.000**	23.9	6.36	**.004**
	Category List	26.8	18.75		41.0	13.8	

TABLE VII. POWER ANALYSIS

Dependent variables	**ATM**				**IBS**			
	Observed Power	***Mini. Effect size***	***Effect size***	***P_ value***	***Observed Power***	***Mini. Effect size***	***Effect size***	***P_ value***
Overall Completeness	0.705	0.512	0.432	.144	--	--	--	.009
Overall correctness	0.891	--	0.077	0.822	0.599	0.526	0.376	.105



(a)Overall completeness at High ability (b) Overall completeness at Low ability (c) Overall correctness at High ability

(d) Overall Correctness at Low ability (e) Required Effort at High ability (f) Required Effort at Low ability

Fig. 1. Analysis results using Interaction Plots

A. Discussion

This experimental study investigates the effectiveness of noun phrase technique and category list technique on the quality of domain model. As already mention that the quality of domain model evaluated in terms of completeness, correctness and effort required for constructing the domain model. We summarize the significant results of a main hypothesis.

RQ1: What is the effect of noun phrase and category list technique on quality of the domain model?

This research question consists of number of hypothesis, and each hypothesis represents different domain model elements.

The statistically significant difference is only found between the number of Correct Concepts (CC) and Missing Concepts (MC) identified by noun phrase technique and category list technique when subjects deal with ATM system. In IBS system, statistically significant difference is found between the number of Correct Attribute (CA), Extra Attribute (EA) and Missing Attributes (MA). Those subjects who used noun phrase technique produce d large number of attributes. Some of the attributes are valid. But most of them are useless. This may be the reason that no specific guidelines were provided to extract attributes from requirement specification using noun phrase technique. After identification of noun and noun phrase, subjects skipped to check each and every noun phrase to decide whether it's a concept or attribute. In contrast, using category list subjects identified less but valid attributes.

Regarding Overall completeness, a statistically significant difference is found in IBS system only. But both techniques show a lack of domain model completeness. Subjects produced 29% and 23% complete domain model using noun phrase and category list technique respectively in IBS system. On the other hand, subjects produced almost 53% and 46% complete domain model using noun phrase and category list technique respectively in ATM system. From the combined analysis of both software systems, subjects produced 6% more complete domain model using noun phrase technique as compared to category list.

Regarding overall correctness dependent variables, no statistically significant difference is found in Overall Correctness. On average, subjects produced more extra concepts, relationships and attributes while IBS system using noun phrase technique. So we can say that using noun phrase technique, subjects identified a large number of noun phrases. Some of them were correct and mostly were useless. In addition, satisfactory results were not found while subjects modeled the ATM system. This may be due to the reason that A T M system is common system and easier as compared to IBS system. A little statistically significant difference is found between the overall correctness in IBS system in favor of category list, but due to the low statistical power we cannot reject the null hypothesis about overall correctness.

RQ2: Which domain modeling technique required more effort to design a domain model?

Regarding required effort statistically significant difference is found b e t w e e n both groups to design domain model. Subjects used more time to design domain model using category list technique.

V. Conclusion

There are two basic techniques to model problem domain i.e. noun phrase and category list. In category list technique, Larman [1] provided a list of candidate conceptual classes, which consists of many categories that are important to the business information system. Noun phrase technique is a grammatical analysis of use case description to recognize conceptual classes.

To evaluate the impact of category list and noun phrase technique on the quality of domain model an experiment was designed and conducted. The purpose of experiment was to investigate that which technique produces high quality domain model in terms of completeness, correctness and effort required to design a domain model.

According to the statistical tests results, category list technique produced more correct concepts in both software system but the difference is statistically significant only in ATM system. So, we can conclude that category list technique is best for identifying concepts which are important to the business world. It also avoids unnecessary concepts in the problem domain. Noun phrase technique is better for identifying attributes for concepts. Both techniques performed same in case of relationships. Overall subjects produce 6% more complete domain model using noun phrase technique however the results are statistically significant in IBS system only. There is no significant difference found between two techniques regarding overall correctness. Minimal significant difference is found in case of IBS system therefore due to low statistical power we cannot reject the null hypothesis. It is also observed that for known system, it does not matter which technique you are using. We suggest that the combined use of both techniques will lead to high quality domain model.

As a future direction the same experiment need to be executed in an industrial environment for more realistic results. In which professional developers are used as subjects and the scenario is also realistic instead of an exemplary one.

REFERENCES

[1] C. Larman, Applying UML and Patterns, Prentice-Hall, 2004.

[2] D. Rosenberg, Use Case Driven Object Modeling with UML – A Practical Approach, APress, Berkeley, USA, 2007.

[3] D. R. K. Scott, Applying Use Case Driven Object Modeling with UML: An Annotated e- Commerce Example, Addison Wesley , 2001.

[4] R. Offen, "Domain Understanding Is the Key to Successful System Development," journal of Requirements Engineering , vol. 7, no. 3, pp. 172-175, September 2002.

[5] J. Evermann and Y. Wand, "Toward Formalizing Domain Modeling Semantics in Language Syntax," IEEE Transactions on Software Engineering, vol. 7, no. 1, p. 21 – 37, January 2005.

[6] P. Stevens, Using UML: software engineering with objects and components, Pearson Education, 2006.

[7] T. Yue, L. C. Briand and L. Yvan, "Facilitating the Transition from Use Case Models to Analysis Models: Approach and Experiments," ACM Transactions on Software Engineering and Methodology (TOSEM), vol. 22, no. 1, p. 5, February 2013.

[8] S. Tiwari and A. Gupta, "Does increasing formalism in the use case template help?," in Proceedings of the 7th India Software Engineering Conference, 2014.

[9] S. Tiwari and A. Gupta, "A controlled experiment to assess the effectiveness of eight use case templates.," in 20th Asia- Pacific Software Engineering Conference (APSEC), 2013.

[10] S. Espana, M. Ruiz and A. González, "Systematic derivation of conceptual models from requirements models: a controlled experiment.," in IEEE Sixth International Conference on Research Challenges in Information Science (RCIS), Valencia, 2012.

[11] B. Anda and D. I. Sjøberg, "Investigating the Role of Use Cases in the Construction of Class Diagrams," journal of Empirical Software Engineering, vol. 10, no. 3, p. 285 – 309, July 2005.

[12] L. Briand, Y. Labiche and R. Madrazo-R, "An Experimental Evaluation of the Impact of System Sequence Diagrams and System Operation Contracts on the Quality of the Domain Model," in International Symposium on Empirical Software Engineering and Measurement (ESEM), Banff, 2011.

[13] U. Erra, A. Portnova and G. Scanniello, "Comparing two communication media in use case modeling: results from a controlled experiment.," in ACM- IEEE International Symposium on Empirical Software Engineering and Measurement, 2010.

[14] C. Wohlin, P. Runeson, M. Host, M. C. Ohlsson, B. Regnell and A.Wesslen, Experimentation in Software Engineering, Springer Science & Business Media, 2012.

[15] Poul and o. Jeff, Introduction to software testing, 2008.

[16] Stevens, Perdita and R. J. Pooley, Using UML: software engineering with objects and components., Pearson Education, 2016.

[17] E. R. Giden, ANOVA Repeated Measures, SAGE Publications.

[18] G. Scanniello, C. Gravino and G. Tortora, Does the combined use of class and sequence diagrams improve the source code comprehension?: results from a controlled experiment, 12 Proceedings of the Second Edition of the International Workshop on Experiences and Empirical Studies in Software Modeling ACM, 2012.