IJACSA

WHERE WISDOM SHARES

International Journal of Advanced Computer Science and Applications

Volume 4 Issue 8

August 2013

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

SAI

www.ijacsa.thesai.org

IJACSA

WHERE WISDOM SHARES

INTERNATIONAL JOURNAL OF ADVANCED COMPUTER SCIENCE AND APPLICATIONS

SAI

THE SCIENCE AND INFORMATION ORGANIZATION

www.thesai.org | info@thesai.org

Editorial Preface

From the Desk of Managing Editor…

It is our pleasure to present to you the August 2013 Issue of International Journal of Advanced Computer Science and Applications.

Today, it is incredible to consider that in 1969 men landed on the moon using a computer with a 32-kilobyte memory that was only programmable by the use of punch cards. In 1973, Astronaut Alan Shepherd participated in the first computer "hack" while orbiting the moon in his landing vehicle, as two programmers back on Earth attempted to "hack" into the duplicate computer, to find a way for Shepherd to convince his computer that a catastrophe requiring a mission abort was not happening; the successful hack took 45 minutes to accomplish, and Shepherd went on to hit his golf ball on the moon. Today, the average computer sitting on the desk of a suburban home office has more computing power than the entire U.S. space program that put humans on another world!!

Computer science has affected the human condition in many radical ways. Throughout its history, its developers have striven to make calculation and computation easier, as well as to offer new means by which the other sciences can be advanced. Modern massively-paralleled super-computers help scientists with previously unfeasible problems such as fluid dynamics, complex function convergence, finite element analysis and real-time weather dynamics.

At IJACSA we believe in spreading the subject knowledge with effectiveness in all classes of audience. Nevertheless, the promise of increased engagement requires that we consider how this might be accomplished, delivering up-to-date and authoritative coverage of advanced computer science and applications.

Throughout our archives, new ideas and technologies have been welcomed, carefully critiqued, and discarded or accepted by qualified reviewers and associate editors. Our efforts to improve the quality of the articles published and expand their reach to the interested audience will continue, and these efforts will require critical minds and careful consideration to assess the quality, relevance, and readability of individual articles.

To summarise, the journal has offered its readership thought provoking theoretical, philosophical, and empirical ideas from some of the finest minds worldwide. We thank all our readers for their continued support and goodwill for IJACSA. We will keep you posted on updates about the new programmes launched in collaboration.

Lastly, 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.

We hope that materials contained in this volume will satisfy your expectations and entice you to submit your own contributions in upcoming issues of IJACSA

Thank you for Sharing Wisdom!

Managing Editor
IJACSA
Volume 4 Issue 8 August 2013
ISSN 2156-5570 (Online)
ISSN 2158-107X (Print)

Editorial Board

Reviewer Board Members

CONTENTS

Paper 1: Autonomic Computing for Business Applications
Authors: Devasia Kurian, Pethuru Raj
PAGE 1 – 8

Paper 2: Novel MIME Type and Extension Based Packet Classification Algorithm in WiMAX
Authors: Siddu P. Algur, Niharika Kumar
PAGE 9 – 13

Paper 3: Design and Development of AlgoWBIs
Authors: Kavita, Dr. Abdul Wahid, Dr. G N Purohit
PAGE 14 – 18

Paper 4: Mining Opinion in Online Messages
Authors: Norlela Samsudin, Abdul Razak Hamda, Mazidah Puteh, Mohd Zakree Ahmad Nazri
PAGE 19 – 24

Paper 5: English to Creole and Creole to English Rule Based Machine Translation System
Authors: Sameerchand Pudaruth, Lallesh Sookun, Arvind Kumar Ruchpaul
PAGE 25 – 29

Paper 6: A Hybrid Method to Improve Forecasting Accuracy Utilizing Genetic Algorithm –An Application to the Data of Operating equipment and supplies
Authors: Daisuke Takeyasu, Kazuhiro Takeyasu†
PAGE 30 – 41

Paper 7: User-Based Interaction for Content-Based Image Retrieval by Mining User Navigation Patterns.
Authors: A.Srinagesh, Lavanya Thota, G.P.Saradhi Varma, A.Govardhan
PAGE 42 – 49

Paper 8: A Novel Expert System for Building House Cost Estimation: Design, Implementation, and Evaluation
Authors: Yasser A. Nada
PAGE 50 – 55

Paper 9: On Integrating Mobile Applications into the Digital Forensic Investigative Process
Authors: April Tanner, Soniael Duncan
PAGE 56 – 61

Paper 10: Development of Social Media GIS for Information Exchange between Regions
Authors: Syuji YAMADA, Kayoko YAMAMOTO
PAGE 62 – 73

Paper 11: On The Performance of the Gravitational Search Algorithm
Authors: Taisir Eldos, Rose Al Qasim
PAGE 74 – 78

Paper 12: Evaluation of Human Emotion from Eye Motions
Authors: Vidas Raudonis, Gintaras Dervinis, Andrius Vilkauskas, Agne Paulauskaite - Taraseviciene, Gintare Kersulyte - Raudone
PAGE 79 – 84

Paper 13: Bootstrapping Domain Knowledge Exploration using Conceptual Mapping of Wikipedia
Authors: Mai Eldefrawi, Ahmed Sharaf eldin Ahmed, Adel Elsayed
PAGE 85 – 95

Paper 14: Dynamic Evaluation and Visualisation of the Quality and Reliability of Sensor Data Sources
Authors: Ritaban Dutta, Claire D'Este, Ahsan Morshed, Daniel Smith, Aruneema Das, Jagannath Aryal
PAGE 96 – 103

Paper 15: Virtual Calibration of Cosmic Ray Sensor: Using Supervised Ensemble Machine Learning
Authors: Ritaban Dutta, Claire D'Este
PAGE 104– 110

Paper 16: Debranding in Fantasy Realms: Perceived Marketing Opportunities within the Virtual World
Authors: Kear Andrew, Bown Gerald Robin, Christidi Sofia
PAGE 111 – 117

Paper 17: Cross Language Information Retrieval Model for Discovering WSDL Documents Using Arabic Language Query
Authors: Prof. Dr.Torkey I.Sultan, Dr. Ayman E. Khedr, Fahad Kamal Alsheref
PAGE 118 – 129

Paper 18: Actions for data warehouse success
Authors: Aziza CHAKIR, Hicham MEDROMI, Adil SAYOUTI
PAGE 130 – 133

Paper 19: Intelligent Ambulance Traffic Assistance System
Authors: RONOJOY GHOSH, VIVEK SHAH, HITESH AGARWAL, ASHUTOSH BHUSHAN, PRASUN KANTI GHOSH
PAGE 134 – 141

Paper 20: Investigate the Performance of Document Clustering Approach Based on Association Rules Mining
Authors: Noha Negm, Mohamed Amin, Passent Elkafrawy, Abdel Badeeh M. Salem
PAGE 142 – 151

Paper 21: Directed and Almost-Directed Flow Loops in Real Networks
Authors: M.Todinov
PAGE 152 – 161

Paper 22: Workshop Session Recordings on Green Volunteering Activities of Students in a Disadvantaged Area According to the Good-Hearted Vocation Teacher to Support Itinerant Junk
Authors: Kuntida Thamwipat, Thanakarn Kumphai
PAGE 162 – 166

Paper 23: Knowledge Management Strategyfor SMEs
Authors: Kitimaporn choochote
PAGE 167 – 172

Paper 24: A Proposed NFC Payment Application
Authors: Pardis Pourghomi, Muhammad Qasim Saeed, Gheorghita Ghinea
PAGE 173 – 181

Paper 25: An Analysis of Brand Selection
Authors: Kazuhiro Takeyasu†, Yuki Higuchi

PAGE 182 – 188

Paper 26: Protein PocketViewer: A Web-Service Based Interface for Protein Pocket Extraction and Visualization
Authors: Xiaoyu Zhang, Martin Gordon
PAGE 189 – 193

Paper 27: Generating an Educational Domain Checklist through an Adaptive Framework for Evaluating Educational Systems
Authors: Roobaea S. AlRoobaea, Ali H. Al-Badi, Pam J. Mayhew
PAGE 194 – 200

Paper 28: Semantic, Automatic Image Annotation Based on Multi-Layered Active Contours and Decision Trees
Authors: Joanna Isabelle OLSZEWSKA
PAGE 201 – 208

Paper 29: Representation Modeling Persona by using Ontologies: Vocabulary Persona
Authors: GAOU Salma, EL KADIRI Kamal Eddine, CORNELIU BURAGA Sabin
PAGE 209 – 217

Paper 30: A Study on the Conception of Generic Fuzzy Expert System for Surveillance
Authors: Najar Yousra, Ketata Raouf, Ksouri Mekki
PAGE 218 – 227

Paper 31: The Determination of Affirmative and Negative Intentions for Indirect Speech Acts by a Recommendation Tree
Authors: Takuki Ogawa, Kazuhiro Morita, Masao Fuketa, Jun-ichi Aoe
PAGE 228 – 235

Paper 32: Performance Analysis of Keccak f-[1600]
Authors: Ananya Chowdhury, Utpal Kumar Ray
PAGE 236 – 241

Paper 33: Software Ecosystem: Features, Benefits and Challenges
Authors: J.V. Joshua, D.O. Alao, S.O. Okolie, O. Awodele
PAGE 242 – 247

Paper 34: Improved QO-STBC OFDM System Using Null Interference Elimination
Authors: K. O. O. Anoh, R. A. Abd-Alhameed, Y. A. S. Dama, S. M. R. Jones, T. S. Ghazaany, J. Rodrigues, K. N. Voudouris
PAGE 248 – 253

Paper 35: Pilot Study of Industry Perspective on Requirement Engineering Education: Measurement of Rasch Analysis
Authors: NOR AZLIANA AKMAL JAMALUDIN, SHAMSUL SAHIBUDDIN
PAGE 254 – 259

Paper 36: A Novel Steganography Method for Hiding BW Images into Gray Bitmap Images via k-Modulus Method
Authors: Firas A. Jassim
PAGE 260 – 263

Autonomic Computing for Business Applications

Devasia Kurian
Christ University, Hosur Road,
Bangalore 560029 Karnataka,
India

Pethuru Raj
Senior Consultant Wipro Technologies,
Bangalore 560045 Karnataka,
India

***Abstract*—Autonomic computing, a new deployment technology introduced by IBM a decade ago, to manage the ever increasing complexity of IT systems, has become a part of many large scale deployments today. A lot of inroads have been made by autonomic computing in the areas of networking, data centers, storage, and database management. But few attempts have been exercised in business applications such as ERP, SCM or CRM, and Online Retail. In this paper, we would like to dive deeper to extract and explain where the pioneering autonomic computing paradigm stands today and the varied opportunities and possibilities in this area. A simplistic architecture for deployment of autonomic business applications is introduced and illustrated in this paper. A sample implementation of different management modules from various areas is described in order to invigorate the readers. This should form the basis for a newer and nimbler start, and the ubiquitous application of AC concepts to enable business transformation. This paper represents a solid extension of the paper presented in World Congress on Information and Communication Technologies, 2012[1].**

Keywords—Autonomic Computing Architecture; CRM; Service Oriented Architecture; Multi-Agent Architecture

I. Introduction

Our physical body is intrinsically capable of responding to critical situations, such as an accident or fever in a time-bound manner, to overcome that particular situation. If we are caught in a life or death situation such as face to face with a tiger, the inherent autonomous mechanisms in the body release the right amount of adrenalin, and thereby, keep us alert to face the situation. If any kind of abnormality happens or if any noteworthy changes occur unexpectedly, our highly smart and sophisticated autonomic nervous system anticipates, activates, and adjusts our body functions and parameters. In a second scenario, imagine the activities happening in the body of a person running a marathon. Monitoring of his or her blood pressure, glucose level, body temperature, and sensing of any other body parameters is being carried out by multiple autonomous mechansims. Also, if the body hydration level goes down, appropriate signals have to be dispatched to exhibit the sudden dryness of throat, which in turn prompts the person to drink water. All of these happen autonomously due to the inherent capability that is embedded in our nervous system.

This autonomy characteristic has pervaded every component in our body. Specific tasks are being accomplished by specific body parts, but all these are under the central control of our brain. Correct division and delegation of certain tasks comes handy in achieving the overall autonomy mission. Decision at the top level and execution at the local level is the key differentiator. By doing this task separation in an autonomous way, the central brain can be freed to perform intelligent tasks. Monitoring, management, decision-making, and communication at the central level is the most crucial and critical phenomenon for achieving the desired competencies of the autonomic model.

This art of managing complex situations, by smaller autonomous components and freeing the central system, can be deployed in appliances, networks, services, and applications. They need to be empowered to be self-managing in order to accomplish their assigned functionality and responsibility under all the circumstances. They ought to be equipped to self-diagnose, self-configure, self-heal, self-optimize, and self-protect. With IT autonomy set to grow, the goals of IT agility, availability, and affordability will see a neat and nice realization. The utilization and sharing of various IT resources goes up sharply with the reduction of human intervention, interpretation, and instruction. Ultimately, IT dependability can be guaranteed effortlessly, at a reduced monitoring, maintenance, and management cost.

As IT is getting tightly coupled and synchronized to business goals that are frequently changing, the next-generation IT systems and services need to be produced and deployed as viable and valuable autonomic artifacts so that they behave differently from traditional systems. This helps in moderating the rising multiplicity- and heterogeneity-induced complexity. This is the gist and essence of this new emerging and evolving computing paradigm. Analogous to the animal world, the central brain needs to be freed from mundane tasks. If we take the example of a router in a networking environment, the regular analysis of packet heads and subsequent routing can be taken over by a low level task autonomously. The central system can occupy itself with QoS (Quality of Service) analysis of routes, and change in routes.

The trend of applications governed by business goals is getting accelerated due to numerous and notable advances in in computing, communication, connectivity, collaboration, sensing, perception, and actuation technologies. Right and relevant innovations and inclinations in IT infrastructures and the smart leverage of analytics go a long way in shaping up the business IT.

They help in visualizing all kinds of barriers and to overcome them proactively and preemptively with competent technological improvisations. In this context, the emergence and evolution of autonomic computing is a real boon and blessing for the struggling IT industry. The introduction of autonomic computing in business applications will bring in a disruptive change in this area. In a similar manner as the consistent evolution of natural elements, it will help in the evolution of flexible and futuristic business applications.

II. SURVEY OF AUTONOMIC JOURNEY

In 2001, IBM had internally initiated and investigated the ground-breaking autonomic idea. Through its seminal paper in 2003 [2], IBM had articulated its vision, design principles, development methodologies, and prospects to the outside world. The basic concepts of the MAPE-loop (Monitor-Analyze-Plan-Execute) as fundamental characteristics of an autonomic component and the major self-CHOP (self-configuration, -healing, -optimisation, -protection) characteristics were defined.

As IT systems become more and more complex, there is a need being felt for powerful and cutting-edge technologies and solutions in order to radically minimize and moderate the heterogeneity and multiplicity-induced IT complexity. Professors and professionals have been cooperating to cultivate and inculcate the innovation spirit among students and scholars to bring forth a bevy of competent solutions for the ills afflicting autonomic computing.

Transformational and Optimization Methods - Control systems theory is being leveraged to provide a theoretical and mathematical framework [3] to analyze, in depth, the behavior of autonomic systems and to accurately predict or control parameters such as stability, settling times, and accurate regulation. A variant of differential evolution [4] as the mathematical basis for optimization questions is emerging. Policy frameworks and engines have been established to manage the systems at higher levels of abstraction [5]. Attempts are being made to attain a refinement of policy frameworks using Ontologies [6] and advanced Organic Computing [7]. There are several other theories, techniques, and toolkits to bring in remarkable innovations and improvisations to realize competent autonomic systems of all sizes and scopes.

Domain Penetration – Every kind of electronic systems, especially IT systems, enabling business automation today need well-intended autonomy. Networking and communication fields are prominent and dominant in assimilating and articulating the trailblazing autonomic concepts. A number of research efforts have been initiated and are being sustained to have the brilliant scintillating autonomic features into telecom systems. Network connectivity solutions such as switches, routers, gateways, and proxies are being embedded with autonomy capabilities to behave adaptively.

Robotics is another interesting and inspiring area where the autonomic principles are greatly recognized and rewarded. Humanoids are the latest incarnation of cognitive robotic systems that imitate humans not only in their structures, but also in their behaviors.

Autonomic Solutions - Real-time performance engineering and enhancement (PE2) of mission-critical business systems and server machines can be achieved through the maturity and stability of autonomic concepts. For example, all types of server machines such as web, application, and database systems are being spruced up to ensure high performance and throughput through a host of optimization and automation mechanisms [8]. There are scholarly projects trying to converge the fine and exemplary facets of autonomic computing principles with the fast-growing cloud computing. It is very clear that a number of server and management tasks (user load, workload, job scheduling, provisioning and de-provisioning of various IT resources) are getting automated through software solutions [9].

A few sample projects deployed in the time period starting from 2003 are:

AUTONOMIA [10] presents one of the initial architectures implementing self-configuring and self-healing characteristics. It introduces an Autonomic Middleware service (AMS) comprising of a component and resource repository, and Fault and Security Handlers.

FOCALE [11] introduces a semantically rich architecture for orchestrating the behavior of heterogeneous and distributed computing elements with support of ontologies, policies and knowledge engineering.

PAWS [12] presents an adaptive framework based on web-services. A BPEL (Business Process Execution Language) editor is provided, with which the business process and the constraints in terms of QoS can be defined.

SASSY [13] provides a framework for systems adapting to requirements by selecting services from service providers, that satisfy the utility function. SASSY introduces SAS (Service Activity Schemas), a visual requirements specification language, which describes the required services and activities from the domain ontology, and SSS (Service Sequence Scenarios), which defines the OS requirements. SASSY works in the framework of an SOA architecture to provide the self-architecting framework.

IPAutomata [14] is a part of IPCenter, IPSoft’s commercial product for enterprise wide service delivery, an autonomic component managing multiple intelligent agents.

Current Research Focus – During the past decade, autonomous computing has come out of its infancy, and is slowly becoming a mainstream technology in the areas of networks, data centers, storage, database management and so on. The research focus being discussed currently [15] are interoperability issues in heterogeneous environments, certification, standardization, and so on.

III. AUTONOMIC COMPUTING – CHALLENGES

Despite all efforts and progress made (as listed above) the autonomic computing discipline has still remained a low-key technology and has not gained as much prominence as other computing models such as cloud computing.

The design methodology for autonomic applications has not grown as much as the popular enterprise application frameworks such as the Microsoft .NET framework or the matured Java Enterprise Edition (JEE). The prototype implementations made so far, by various companies across the globe, have miserably failed to ignite widespread technology awareness. Also, the commercial successes achieved of deployed autonomic systems have not raised the pitch for autonomic computing. A thorough understanding of the possible reasons could help us work on these in the future and bring in the deserved importance and desired results to this transformative, disruptive, and innovative technology.

- **Industry Support –** IBM has laid the foundation for the autonomic approach to sharply minimize the complexity of IT systems and has invested a lot in terms of people, money, and so on, in order to take it forward. Unfortunately, IBM has failed to move it into the IT industry in a big way. It has failed to build a larger consortium for formulating standard specifications and to popularize a simplified and streamlined design methodology within the software development community.

- **Less Awareness -** Unfortunately, Autonomic Computing has not been able to penetrate the ecosystem and get wide spread acceptance. It is being promoted as a specialized niche technology. Its visibility and vitality factors are not appropriately disseminated to the IT market.

- **System Characteristics –** The seminal paper from IBM [2] established the self-managing characteristics as fundamental for autonomous systems. Many systems tried to incorporate one or more characteristics rather than focusing on a single aspect. This led to a dilution of resources, which ultimately resulted in weak systems with zero or minimal commercial impact. Furthermore, many of the projects introduced new characteristics like self-management.This resulted in dilution of resources.

- **The Positioning Conundrum –** The ever increasing complexity of IT systems can be solved at three different levels of sophistication as illustrated in the diagram below.

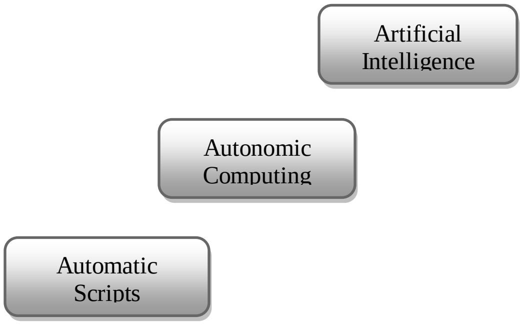


Fig. 1. Autonomic Computing Positioning

Most of the implementers are not clear about this distinction: scripts to carry out certain activities, autonomic systems to manage activities with a feedback loop or artificial intelligence to cognitively solve the issues in the system. During the implementation, they tend to give way to easier scriptsor dig their heads in the myriads of artificial intelligence. The purist view of architecting autonomic computing systems as a simple control system element is forgotten, and the studies go in depth for issues like dynamic location of input services, knowledge transfer, ontologies and so on [16]. This results in researchers spending most of the time fixing semantic or knowledge issues, and thereby forget to exploit the power of the autonomous control loop. Focus should be brought back to the implementation of autonomic computing systems with fixed set of inputs, control function, and control variables.

- **Generic Architecture -** Initial approaches to design autonomous systems came from various corners like control systems, agent-based systems, component-based systems, and so on. The interface mechanism for any agent-based system varied from the interface mechanism for component-based systems. Multiple deviating approaches have destroyed the emergence of a generic architecture and interfaces for autonomous systems.

- **Lack of Standards –** Standards generally inspire worldwide product vendors to produce best-of-the-breed implementations. Standards facilitate interoperability among diverse and distributed system modules. Technologies such as JEE, which have been dominating the technology scenario, are implicitly supported and sustained with simple and pragmatic standards.

- **Lack of Toolkits -** Besides standards, facilitating frameworks, enabling toolkits and platforms, best practices, design patterns, optimized processes, and key guidelines also play a very important role in the massive adoption of technology. Other than ATK (Autonomic Computing Toolkit), a part of the ETTK (Emerging technologies tool kit), produced by IBM, there have not been many simple and sensitive toolkits for AC.

These are few barriers for autonomic computing that responsible for its inability to catch up with its peers. The above issues have to be kept in mind while formulating a framework for business applications. A thorough understanding of the above issues will help one to formulate the right framework for autonomic computing in business applications.

IV. Autonomic Business Applications Architecture

Typically, an autonomic system should have multiple independent, yet connected, autonomic components providing a variety of services. The services could be monitoring, providing policy rules, decision-making, controlling, managing, or actuation services. Intelligent software agents are the prime building blocks for autonomic systems and hence they are called multi-agent systems. The individual agents can be components or services implementing a particular function or entities providing multiple services [17]. Policy or rule based engines are the other prominent contributors for systems to be autonomic in their dealings and deliveries. The knowledge manager is a kind of resource manager facilitating insightful access and retrieval of all kinds of information from the knowledge bases, so that appropriate decisions can be contemplated and conveyed.

In short, it is about empowering and embodying software with extra qualities and capabilities. This can be done by employing specially designed engines and knowledge bases so that the software can adaptively exhibit human-like behavior in discharging and delivering their assignments. We have crafted macro-level architecture for an open and standards-based framework for quickly realizing adaptive and self-evolving software. The proposed framework as depicted in Figure 2 will have the following:

1) *Autonomic Managers*
2) *Knowledge Base*
3) *Enterprise Service Bus (ESB)*
4) *Policy Framework and Language*
5) *Reusable assets repository for components (RAR)*

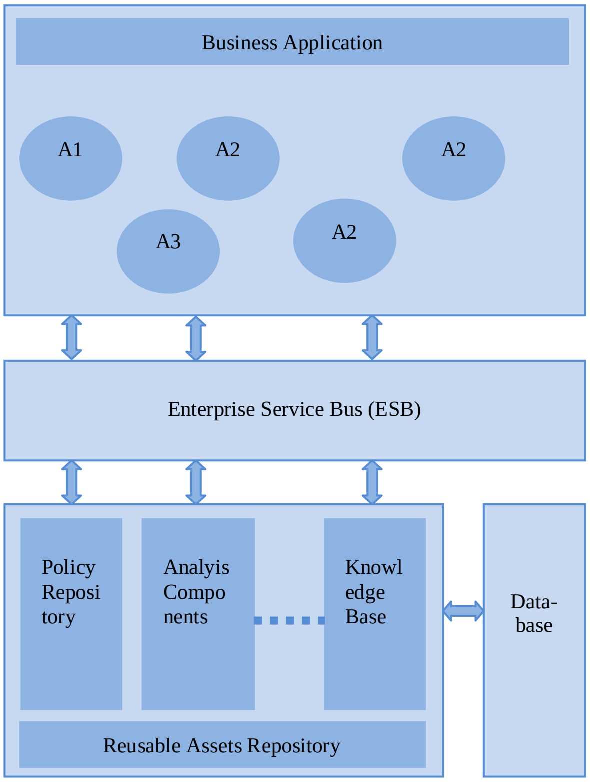


Fig. 2. Autonomic Computing Architecture

The autonomic managers will run as independent processes following a multi-agent architecture. The managers access relevant policies from the repositories and carry out actions based on the specifications in the policies. The evaluation of a policy and its relevant data happens through multiple services provided by various components or agents in the system. To facilitate a dynamic, flexible, and secure access to these services, an ESB framework will be used.

The RAR component will help to maintain and manage components and services in an efficient manner. For small implementations, the ESB and RAR component could be opted out. In such cases, the autonomous managers will directly interact with the component services.

Designing applications as autonomic components and remodeling existing applications as autonomic components are two ways to incorporate autonomy in systems. In the latter case, the vendors should provide service interfaces to monitor or manipulate application data so that third party vendors can create third party autonomic managers. This open approach will help specialists in an area to create relevant autonomic managers, for example a data mining company creating an autonomic manager to find associations using superior associative algorithms, or a text analysis company creating managers to extract key words from data.

Many times cost and time factors are an impediment to creating a system with autonomic managers from scratch. In such cases, existing systems can be extended by autonomic components directly manipulating the database. This is not a recommended practice because there is uncontrolled manipulation of data from multiple sources and therefore, such developments need to be tested thoroughly.

V. Sample Autonomic Computing Applications

The simplicity and application of individual managers is best explained with industry relevant examples. Autonomic managers could be introduced into existing applications, which could control activities of individual modules. It is assumed that the developer of original ERP is not involved in these introductions and therefore, we are dependent on direct database manipulations. If the original developer is the initiator of the introduction of autonomic components, then this could be carried out more elegantly in one of the following ways:

- **Design level:** Autonomic behavior is built in during the design time itself.
- **Interface Level:** At the design level, interfaces are defined to enable other developers to manipulate certain parameters in a controlled manner. This might not be only parameter manipulations, but also interfaces to influence the system status like start, stop, and so on.

The idea of introducing the examples is to:

- Illustrate the options of developing autonomic components for existing applications,
- Initiate the evolution of similar ideas among the domain practitioners and develop standard definitions, interfaces, and so on for independent autonomic component developers.

This would create an ecosystem of core application developers and intelligent peripheral autonomic component developers.

A. *Sample Autonomic Computing for CRM*

At the outset, CRM is a typical monolithic applicationwhere the introduction of autonomous components might not seem appropriate. In most cases, customers using traditional CRM systems slowly start sinking in the huge volume of data produced by these systems. The introduction of autonomic components into this monolithic architecture brings better manageability and quicker approach to their business goals.

The following sample autonomic computing managers will help the users of CRM systems manage the complexities that are increasing day by day:

Lead Generation Manager: A primary function of any CRM is the management of leads, including all details and communication. The Lead Generation Manager is a self-optimizing autonomic component, which goes into a loop as depicted in Figure 3.

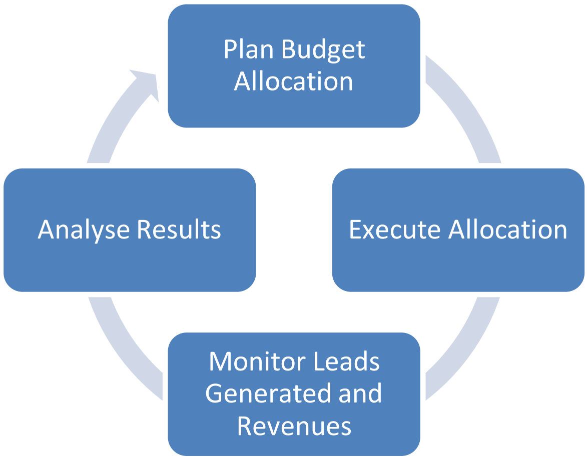


Fig.3. Lead Generation Manager

The administrator specifies in the policy manager that the lead generation should be optimized for a utility function, for example, to maximize the revenue generated by the leads. The optimization could have be carried out based on the number of leads generated, proposals sent, and so on. The budget for lead generation activities is one of the parameters for the policy. The plan component allocates the budget to various lead generation activities – a) online: such as search engines and web-site banners and b) offline: such as exhibitions and print advertisements. The execution takes place automatically, as in the case of many online promotions or manually, as in the case of exhibitions. The revenue generation results are assimilated with pointers to the lead generation activity. This information is analyzed to find out the most effective lead generation media and the budget is reallocated to each of these media based on rules specified in the policy. A simple greedy algorithm might suffice in most of the cases.

PricingManager: A pricing manager could optimize pricing policies across various pricing schemes. The pricing manager may deploy various pricing schemes from season to season and measure the effectiveness of each in terms of revenues, profits, and so on. Based on suggested algorithms, efficient schemes can be selected.

ReminderManager: Customer relationships are maintained by effective reminders of various activities throughout the life-cycle like callback after initial contact, reminder after proposal, activation after a period of inactivity and so on. The optimal results of the reminders differ for each customer segment and product, based on elapsed time, type of reminder, and so on. For example, a customer, who is interested in a buying children health drink might revisit the idea either after 1 week (if she has not yet bought it) or after 6 weeks (to buy again, after finishing the competitor product she bought). The ReminderManager activates various reminders for products, measures the effectiveness, and optimizes the reminder process for the given customer segment and product.

B. *Sample Autonomic Computing for ERP*

Enterprise Resource Planning systems today encompass multiple modules taking care of planning, tracking, and controlling multiple resources and activities in an enterprises. The systems have become so complex, that each enterprise employs a team of engineers to ensure the proper running of the system. This results in huge expenses every year. Often, ERP becomes so complex that it is a mere collection of various modules and interactions between these modules are not regulated. These modules could be production planning, purchasing, inventory, finance, HR, and even extended modules like CRM and so on.

Some of the typical managers, which could be introduced in an existing ERP system are:

InventoryManager: Inventory management is a very important section in ERP, which has great implications on the financial performance of the company. Most of the ERP systems implement simple parameter watches minimum stock quantity, maximum stock quantity, minimum order quantity and so on.

The autonomic Inventory management module collects data from multiple modules such as production, sales, and purchase, executes daily purchase parameters, and measures the effects. The main utility function is the inventory turnover ratio. Sub-parameters such as average inventory cost, and average wait time might also influence the purchase pattern.

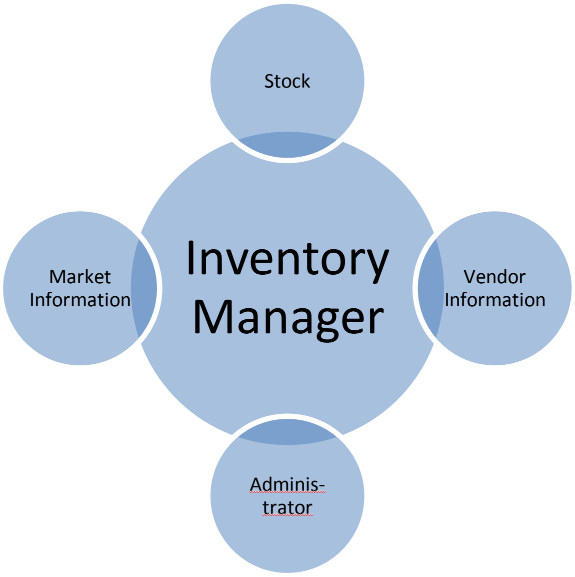


Fig.4. Inventory Manager

Administrator feedback might be sought for special cases, where the autonomic manager might not be able to take a direct decision, for example, when the price of a material drops suddenly. In this case, a lot more information available external to the system is required to decide if it is advisable to procure large quantities for later usage. A query is sent to the administrator giving specifics of the situation and actions are taken based on the recommendation of the administrator.

VendorManager: Long term Vendor Management with selection and rating is very important to the supply side of any organization. Normally, companies keep two or more vendors for each product and divide the purchases, in a manner to optimize the supply, and at the same time keep all of them interested in supplying.

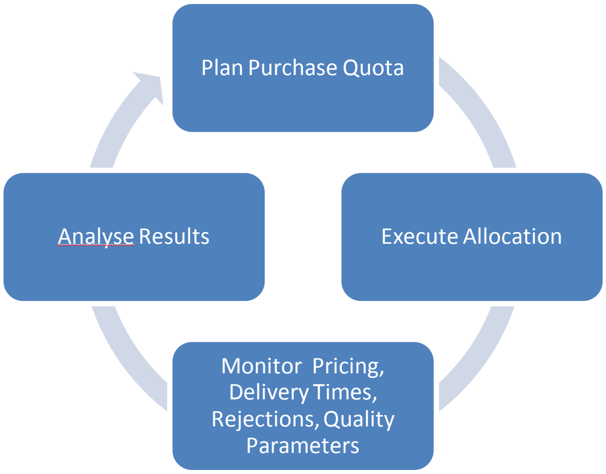

Plan Purchase Quota
Execute Allocation
Monitor Pricing, Delivery Times, Rejections, Quality Parameters
Analyse Results

Fig.5. Vendor Management

The vendors are rated on various feedback parameters such as pricing, delivery times, quality, and so on. Each time a purchase has to be made, the purchase quantities are allocated based on the ratings.

Similar managers can be deployed in multiple areas of ERP such as production, finances, HR, CRM, and SCM. The autonomic integration framework ensures the addition of features in a smooth manner without affecting the existing functionality.

VI. Additional Features

The vision of putting together a software system with the help of multiple autonomic components is similar to the process of putting together hardware components to build a PC. IBM started a similar revolution in the world of desktop PCs by opening up the hardware design, components details, and interface details of the PC. For example, the hard disk or mouse is, to a large extent, autonomous in itself with very well defined interfaces. Multiple vendors came in who specialized in individual components enabling specialization and mass manufacturing, which made the PC prices to plummet. The plummeting prices created an upward spiral for demand and quality of the products. A similar revolution is the long term dream of a software engineer to build such systems.

Some salient features of the above system, other than basic autonomous capabilities, are:

Hot-Plugging: Managers, policies, and services can be changed during runtime of the system, without shutting down the complete system.

Extendable: The simple architecture can extend existing application systems for autonomic capabilities, and can also incorporate advanced features based on AI algorithms, semantic technologies, event processing architecture and so on.

Vendor independent: The autonomic managers could be swapped from one vendor to another, if the performance of one vendor is not satisfactory. The idea is that multiple vendors manufacture autonomic components that compete with each other.

New Technology Integration: Technology is evolving in all field. For example, data mining has brought in a complete new set of intelligence to existing applications. The challenge is to incorporate this new knowledge in existing legacy applications. The independent agent framework introduced by autonomic computing provides an excellent platform to integrate such new intelligence with existing systems.

Module inter-dependencies: Advances in technology bring in new forms of interaction between the various modules in a system. For example, the inventory management in an ERP might be influenced by the sales pattern of another product. This type of cross product influence was not thought of during the conception of the ERP system. Autonomic computing architecture is suited to implement such interactions.

Deployment of autonomic computing architecture is also helpful to achieve the above listed features.

VII. The Evolving Technologies for Autonomic Computing

The simple architecture presented in the previous section can be taken forward with the immense contributions from multi-disciplinary researchers and system professionals. There are many untested yet viable and promising technologies existing and emerging for the reality of autonomic systems. The leading ones include artificial intelligence (AI) technologies for vision, perception, natural interfaces, speech recognition, decision-making and actuation, intelligent agents & multi-agent systems. Technologies such as smart components for software-based solutions, semantic technologies, grid computing, storage, exchange, access, and leverage, service-enablement of IT resources, virtualization, analytical solutions, knowledge engineering, organic computing are also promising.

Services will be empowered through semantic technologies such as ontology in order to be semantic-enabled. Semantic services are dynamic in locating other services and using them to create newer composite services that could meet all kinds of requirement changes. This also brings in intelligent processes.

Process optimization, innovation, integration, integrity, and flexibility along lean processes can be realized through this fast emerging dynamic composition phenomenon. Apart from leveraging semantic services, we will develop and install a taxonomy-based knowledge base. Services capable of reasoning and fulfilling situational (context) changes can readily access the knowledge-base to change their behaviors, to make relevant and rightful structural changes, and to embark on appropriate and preferred decision-making activities in real-time.

Enterprises are extremely event-driven these days. A host of business events, such as goods passing through entrances fitted with RFID readers, are the business realities these days. Enterprise systems need to be equipped with relevant software as well as hardware infrastructures in order to cope with these evolutions, so that they are competitive to their customers, clients, retailers, employees, and other stakeholders. Enterprise infrastructures are therefore being strengthened for receiving, consuming, processing, and routing streams of events at real-time besides triggering timely response and counter measures. There are competent event stream processing (ESP) and complex event processing (CEP) solutions in the marketplace.

With standards-compliant toolkits and SOA workbench, a library of services can be quickly realized, stocked, and leveraged. With the advancements being made in Web 3.0 and ontology engineering domains, constructing semantic services will become relatively easier. In short, to realize adaptive and self-evolving software, we need semantic services for automatic discovery, matchmaking, collaboration, and composition and a knowledge base for self-evolvement.

VIII. The Emergence of Autonomic Cloud Center

With the increasing popularity of autonomic concepts, there will be a seamless convergence and consolidation in the form of innocuously embedding the autonomy characteristics with cloud technologies. The manageability of the ever-increasing complexity of cloud infrastructures, platforms, and software can be simplified only by the evolvement and involvement of higher-level concepts such as autonomic computing.

As the expectations on the cloud technology are on the rise, the focus is on smartly leveraging the autonomic concepts in order to make the cloud more relevant for future IT. As cloud is being touted as the next-generation IT environment for hosting and delivering all kinds of IT infrastructures, platforms, and applications such as services over the open and public Internet communication infrastructure, the relevance of the trend-setting autonomic concepts for the cloud is set to grow exponentially.

Autonomic cloud computing is about empowering cloud infrastructures and platforms to take their own decisions in order to continuously accomplish their assigned tasks. Without any kind of intervention, interpretation, and instruction from humans, cloud systems and services need to consistently provide their functionalities and facilities to subscribers. For example, resource management is one well-known job of cloud IT. As clouds are being positioned as the next-generation IT infrastructure for hosting and delivering a number of applications for personal and professional purposes, the importance of and insistence on bringing cloud-enabled business transformation, simplification, and optimization will only grow.

Cloud computing is leading to transformational changes with stringent requirements on performance or throughput, scalability, security, availability, and extensibility. Their run-time management requires realistic algorithms and techniques for sampling, effective measurement, and characterization for dynamic capacity planning, optimal provisioning, user and load prediction. Models are very important in correctly visualizing the end results and accordingly, all kinds of manipulations can be handled in a better manner. There are load prediction algorithms for cloud platforms. Such algorithms, integrated into the autonomic management framework of a cloud platform, can be used to ensure that the SaaS sessions, virtual desktops, or VM pools are autonomically provisioned on demand in an elastic manner. This approach is suitable to support different load decision systems on cloud platforms with highly variable trends in demand, and is characterized by a moderate computational complexity compatible with run-time decisions.

The value and power of autonomic clouds is rising as a result of the application of autonomic techniques to clouds. This amalgamation results in robust, fault tolerant and easy-to-operate clouds. Such autonomic techniques originate from evolutionary and genetic algorithms, multi-objective, and combinational optimization heuristics, artificial neural networks (ANNs), swarm intelligence and multi-agents systems. These proven and promising techniques can improve the way in which computing systems and applications are built, used, managed, and optimized. Therefore, the benefits for users can be maximized by reducing the operational, maintenance, and usage costs of clouds.

Thus autonomic computing techniques, technologies, and tips are bound to bring in a series of innovations especially state-of-the-art IT infrastructures, platforms, and applications for future IT.

The simple architecture described in the paper also takes care of the deployment of these components in a cloud environment by strictly having multi-agents and SOA for delivering various services.

IX. Future Work

Establishment of a simple architecture for business applications is the requirement of the day to manage the ever increasing complexity of business applications. The industry players should work together to achieve this goal in each domain and create the required definitions, toolkits, and so on to enable multiple vendors to work together in a heterogeneous environment. Some pointers in this direction are:

- **Domain specific Autonomic Definitions:** Vendors of domain specific applications could join hands together to develop specifications of domain control definitions to allow third party vendors to develop autonomic modules. This could be similar to CRM vendors joining hands to define a lead management interface, which could input leads into the system and measure the effectiveness of the leads and thereby, control the lead management campaign. This would lead to a breakdown of a large monolithic application into a collection of smaller components.

- **Dynamic Autonomic Managers:** Vendors can create separate managers that can be integrated into existing systems, such as the Lead Manager for CRM applications. The interface points could be defined at database level or the vendors could create plugins for the popular CRM applications.

- **Analysis Components:** Statistical Analysis or Business Analytics players could provide analysis tools as a service with interface definitions.

- **APIs for Applications:** Application vendors could provide APIs for monitoring and manipulating application specific parameters. For example, a discount interface could allow external applications to manipulate the discount percentages for products available in a system.

Industry forums should come together and actively contribute towards creating definitions of an independent environment to enable the long-term vision of autonomic computing and thereby creating powerful applications. Vendors of specific capabilities such as business analytics should create components and managers in line with the definitions. Similarly, vendors of large applications such as SalesForce, SugarCRM, SAP, and Baan should provide APIs for smaller vendors to interface smoothly with their applications. This ecosystem of all players will push the autonomic computing technology to new heights.

X. Conclusions

IT systems, solutions, and services are becoming complicated with the addition of more and more features and functionalities. As IT complexity is on the rise, there is an urgent need for easy-to-learn and use complexity-mitigation techniques and tools. The much-discussed and discoursed autonomic computing is being prescribed as the most promising and potential mechanism to significantly slash the unbridled growth of IT complexity. We have listed out the ways and means for taking forward this consolidated and comprehensive method to boost the required IT transformation that in turn will have a telling influence on business agility.

Study of existing implementations has revealed that the deployment of autonomic computing techniques has made some progress in networking, load balancing, power management, and so on, but few inroads have been made into business applications. We have introduced a simplified architecture in this paper, keeping in mind the possibilities of wide spread acceptance from multiple technology corners, and the core trends like policies, knowledge segregation, services, and cloud. The proposed implementation of autonomic agents in the areas of ERP and CRM shows that the technology is well suited to implement complex and sophisticated systems of the future.

References

[1] D. Kurian and P. R. Chelliah, "An Autonomic Computing Architecture for Business Applications," in IEEEXplore Digital Library, Trivandrum, 2012.

[2] J. O. Kephart and D. M. Chess, "The Vision of Autonomic Computing," IEEE Computer Society, pp. 41-50, January 2003.

[3] Y. Diao, L. J. Hellerstein, S. Parekh and Griffith, "A Control Theory Foundation for Self-Managing Computing Systems," IEEE Journal on Selected Areas in Communication, vol. 23, pp. 2213--2221, 2003.

[4] R. Storn and K. Price, "Differential Evolution – A Simple and Efficient Heuristic for Global Optimization over Continuous Spaces," Journal of Global Optmization, vol. 11, pp. 341-359, 1997.

[5] R. Calinescu, "Challenges and Best Practices in Policy-Based Autonomic Architectures," in Third IEEE International Symposium on Dependable, Autonomic and Secure Computing, 2007.

[6] L. Stojanovic, J. Schneider, A. Maedche and S. Libischer, "The role of ontologies in autonomic computing systems," IBM Systems Journal, vol. 43, pp. 598-616, 2004.

[7] H. Schmeck, C. Müller-Schloer, E. Cakar, M. Mnif and U. Richter, "Adaptivity and Self-Organisation in Organic Computing Systems," ACM Transactions on Autonomous and Adaptive Systems, vol. 5, no. 10, pp. 1-32, September 2010.

[8] B. Raza, A. Mateen, T. Hussain and M. M. Awais, "Autonomic Success in Database Management Systems," in ACIS International Conference on Computer and Information Science, 2008.

[9] M. Rak and A. Cuomo, "CHASE: an Autonomic Service Engine for Cloud Environments," in 20th IEEE International Workshop on Enabling Technologies: Infrastructure for Collaborative Enterprises, 2011.

[10] X. Dong, S. Hariri and L. Xue, "AUTONOMIA: An Autonomic Computing Environment," 2003.

[11] J. C. Strassner, N. Agoulmine and E. Lehtihet, "FOCALE – A Novel Autonomic Networking Architecture," 2006.

[12] D. Ardagna, M. Comuzzi and E. Mussi, "PAWS: A Framework for Executing Adaptive Web-Service Processes," 2007.

[13] A. D. Menascé, H. Gomaa, S. Malek and P. J. Sousa, "SASSY: A Framework for Self-Architecting," IEEE Software, pp. 78-85, November 2011.

[14] February 2012. [Online]. Available: http://www.ipsoft.com/.

[15] T. O. Eze, R. J. Anthony, C. Walshaw and A. Soper, "Autonomic Computing in the First Decade: Trends and Direction," in IARIA, St. Marteen, 2012.

[16] R. M. Nami and K. Bertels, "A Survey of Autonomic Computing," Athens, 2007.

[17] L. Zhen and M. Parashar, "Rudder: An agent-based infrastructure for Autonomic Composition of Grid Applications," 2005.

Novel MIME Type and Extension Based Packet Classification Algorithm in WiMAX

Siddu P. Algur
Departmentof Computer Science
Rani Chennamma University Belgaum, India

Niharika Kumar
Department of Information Science and Engineering
RNSIT Bangalore, India

***Abstract*—IEEE 802.16 provides quality of service by providing five different service classes. When a packet reaches the MAC layer, the packet classifier has to classify the packet such that the packet is associated with appropriate QoS. In this paper a packet classification algorithm is proposed that exploits the HTTP header of the application layer to determine the type of data and classify the data. The extension and MIME type (content type) headers are used to classify the packets. Further the algorithm is enhanced by considering the type of user as an additional parameter during packet classification. Simulations are done on variable bit rate video data. Based on the type of video the packets are classified as RTPS or nRTPS thereby providing differentiated QoS. Simulation results reveal that high priority video classified as RTPS traffic receive higher throughput compared to video of lower priority. The delay faced by high priority videos is correspondingly less compared to low priority video.**

Keywords—QoS; WiMAX; MAC; Packet Classification; IEEE802.16e

I. INTRODUCTION

IEEE 802.16 also called as WiMAX provides last mile broadband wireless service. WIMAX was initially proposed as a broadband solution for fixed wireless devices. IEEE 802.16e adds mobility support to WiMAX. WiMAX supports broadband speed up to 100 Mbps within a service cell radius of about 5 KM.

WiMAX supports quality of service by providing five different service classes.

The first service class called Unsolicited Grant Services (UGS) is designed to support real time data streams that generate fixed size packets at periodic intervals. Voice over IP without silence suppression is an example for traffic that is categorized as UGS. The second service class called Real Time Polling Services (RTPS) supports real time data streams that generate variable sized packets on periodic basis. For example an MPEG video consists of P Frame, B Frame and I Frame. The third service class also called as Extended Real Time Polling Services (eRTPS) supports real-time service flows that generate variable sized data packets at periodic intervals, for example VoIP with silence suppression. The fourth service class is called as Non Real Time Polling Services (nRTPS). nRTPS supports delay tolerant data streams that generate variable size data packets. An example for one such type of traffic is the file transfer protocol data (FTP). The last service class, also called as Best Effort (BE), supports data streams which do not require any service level. Ex Web browsing, Email etc.

When a packet is generated by the user it is encapsulated in the TCP/UDP packet which is further enclosed within an IP datagram before it reaches the data link layer. The packet classifier at the data link layer classifies the packet to one of five service classes. WiMAX does not specify any packet classification algorithm. Authors have proposed various packet classification algorithms. In [4] IP packet provides certain bits that can be used by packet classification algorithm. In [5] an IP address based packet classification mechanism is proposed. In this paper an extension and MIME type based packet classification for WiMAX is proposed. These methods exploit the HTTP protocol headers for packet classification. As per available literature an application layer based cross layer packet classification has not been studied before.

This paper is divided into following sections. Section II describes the existing packet classification algorithms. Section III proposes “extension” and “MIME type” based packet classification algorithms. Further a user based improvisation to the algorithm is suggested. Section IV provides a model for the system. Section V delves the simulation and a discussion on the simulation results is undertaken. Section VI concludes the paper.

II. EXISTING PACKET CLASSIFICATION ALGORITHM

IEEE 802.16e does not specify any packet classification algorithm. It is upto the vendor to implement an appropriate packet classification mechanism. One method of packet classification uses the type of service (TOS) field of the IP header. TOS field has 3 bits named delay, throughput and reliability. A value of 0 for the three bits indicates normal delay, normal throughput and normal reliability. A value of 1 for the bits indicates low delay, high throughput and high reliability. MAC layer can usethe value of these fields to classify the packets into one of the five service classes. Though this technique provides a method of distinguishing and classifying data into different service classes, it does not provide fine grained support for packet identification and classification. [5] Provides an IP address based packet classification mechanism where the packet classification is performed based on the source and destination IP address. This method of packet classification necessitates the knowledge of ip address of the receiver to result in appropriate packet classification. This is a look up table based packet classification relevant for routers.

III. Proposed Packet Classification Algorithm

Data packets enclosed and transmitted using http protocol have HTTP headers to identify the data. HTTP protocol specifies different type of headers to aid the sender and receiver in processing the data. Two headers that are of interest are the “GET” header in the HTTP request and the “Content-Type” header in the response header.

The packet classifier at the MAC layer in WiMAX can use the GET header and the “Content-type” header to classify the packets at the MS and BS as described below.

A. Extension Based Packet Classification

When an MS issues a HTTP GET request for data, it provides the URL of the data. The url contains the location of the data on the server. The URL ends with the type of content being requested. For example, in case of video packets the URL could end with “.mpeg”, “.mp4”, “.avi” etc. The packet classifier algorithm is described below:

Step 1: Extract the HTTP header from the packet.

Step 2: Extract the GET header from the HTTP headers.

Step 3: Identify the type of request based on the extension (ex: mp4, mpeg etc).

Step 4: Extract the source and destination port number and IP address from the packet.

Step 5: Assign a connection ID representing the flow as RTPS, nRTPS, eRTPS or BE.

Step 6: create a tuple containing six parameters as below:

<extension, connection ID, source IP, destination IP, source port, destination port>

Step 7: place the packet in the appropriate queue (i.e RTPS, nRTPS, eRTPS or BE).

The above packet classification shall be done for the first packet of the service flow. Subsequent packets of the service flow may not contain the get header. So, when subsequent packets arrive they shall contain the source IP, destination IP, source port and destination port. These four parameters in the packet shall be compared with the four parameters of the tuple and when matched the corresponding connection ID is retrieved. This connection ID represents the service flow to which the packet shall be forwarded (RTPS, nRTPS, eRTPS or BE). Hence the packet classification algorithm is able to classify the packets based on the extension of the request.

B. Content Type or MIME type based packet classification.

The response for an HTTP request contains the content requested by the user. The response packets also contain the HTTP headers. Along with the different headers like Cache Control, cookies, Age, Date, the server sends the Content type of the data. For example, if the video is of type avi the content type shall be video/x-msvideo. For MPEG video the mime type shall be video/mpeg and for quick time videos the mime type is video/quicktime. When the packets reach the base station (BS), BS can decode the MIME type to identify the content type and classify the packets accordingly. The algorithm for packet classification based on the MIME type is as below:

Step 1: Extract the HTTP header from the response packet.

Step 2: Extract the Content type / MIME type header from the HTTP headers.

Step 3: Identify the type of content based on the content type/ MIME type (video/mpeg, video/quicktime, video/x-msvideoetc).

Step 4: Extract the source and destination port number and IP address from the packet.

Step 5: Assign a connection ID representing the flow as RTPS, nRTPS, eRTPS or BE.

Step 6: create a tuple containing six parameters as below:

<MIME type, connection ID, source IP, destination IP, souce port, destination port>

Step 7: place the packet in the appropriate queue (i.e RTPS, nRTPS, eRTPS or BE).

This classification and tuple generation is done for the first response packet. For all subsequent response packets the source IP, destination IP, source port and destination port details from the packet is mapped with all the tuples. The matching tuple is extracted and the connection ID is obtained from the tuple. The packet is then forwarded to the queue corresponding to the connection ID (i.e RTPS, eRTPS, nRTPS or BE). Hence the packet gets classified as one of the service class packet based on the MIME type of the packet.

1) Advantage

MIME type and extension based classification provides the advantage of classifying the packets based on the extension and MIME type there by giving a finer control at packet classification. In the general packet classification mechanism all video packets are classified as RTPS traffic. By using the MIME type based packet classification, specific videos can be classified as high priority RTPS traffic and other videos can be classified as low priority nRTPS or BE packets. This lets the network operators to selectively degrade the quality of service for certain types of video traffic during dynamic network conditions. When the network is heavily loaded the high bandwidth MPEG packets can be downgraded to nRTPS/BE and bandwidth efficient video packets like mp4 can be treated as RTPS. This lets the BS serve many users at the same time. When the network condition improves, the MPEG videos can again be classified as RTPS packets to improve the QoS for MPEG users.

C. User oriented MIME type and extension based packet classification

MIME type and extension based bandwidth allocation can be further specialized to selectively classify the same MIME type data differently for different types of users,i.e. for certain users the MPEG videos can be classified as RTPS packets and for some other users the MPEG videos can be classified as nRTPS data. This lets network operators provide graded quality of service based on the type of user and the type of traffic generated for/by the user.

The network could grade the users as priority users and regular users. Priority users could be users who pay more or utilize the network more often. Such users generate more revenue and also ensure a higher utilization factor for the network operators. Such users can be classified as priority users. Regular users could be the users who pay the regular network utilization fee.

The operator can maintain a mapping table of the MAC address of the user and the priority value associated with the user.

So, when a video/mpeg packet is being classified as RTPS, nRTPS or BE packet the BS uses the following algorithm to classify the packet. The request packet obtained from MS and the response packet being sent to MS together contribute to the packet classification algorithm. This two stage process is described below.

1) Stage 1: BS uses extension type and user priority for tuple creation

Step 1: Receive the HTTP request from MS.

Step 2: Extract the MAC address of user from MAC header

Step 3: From the priority table extract the priority associated with user (regular user or priority user)

Step 4: Extract the extension from the GET request to identify the type of data.

Step 5: Based on the extension type and the priority of the user, associate the request to a particular connection. Ex: high priority user generating MPEG packet is associated with RTPS connection and regular user generating MPEG packet is associated with nRTPS/BE connection.

Step 6: Generate a tuple containing seven parameters. One of the parameters (content type is null for now). The tuple looks as below:

<extension, content_type_null, connection ID, source IP, destination IP, souce port, destination port>

2) Stage 2: BS uses the tuple for packet classification

The GET request is then forwarded to the network by the BS. When a response from the network arrives, the BS shall extract the source IP, destination IP, source port, destination port to identify the tuple. The tuple is then updated with the content type to form the complete tuple.

<extension, content_type, connection ID, source IP, destination IP, souce port, destination port>

From the tuple extract the connection ID and place the packet in the queue corresponding to the connection ID (RTPS, nRTPS, eRTPS, BE). Hence a combination of extension content type and the type of user is used to classify the packet appropriately.

IV. System Modeling

Let the system contain users that generate video traffic of different types. Some users generate lossless high resolution MPEG video and other users generate lossy mp4 video. The amount of data generated per video frame for an MPEG video shall be much higher compared to a MP4 video. The packet classifier shall read the content type of the data to determine if it is video/mpeg or video/mp4.

All packets of video/mpeg are assigned to the connection associated with the QoS for nRTPS traffic and all packets of video/mp4 are assigned to the connection associated with the QoS RTPS traffic.

As the MPEG packets keep arriving the packets are queued in the nRTPS queue and all mp4 packets get queued as RTPS traffic.

The base station allocates unicast request opportunities to RTPS queue and unicast/broadcast request opportunities to nRTPS queues to request for bandwidth. Each of the RTPS and nRTPS queues request for bandwidth by sending bandwidth request headers.

When the bandwidth request header reaches BS the base station allocates bandwidth by scanning through the bandwidth request headers. BS shall allocate bandwidth in the following order:

UGS >> RTPS >>eRTPS>> NRTPS >>BE

Since the RTPS connections are allocated bandwidth first, MP4 videos get higher share of bandwidth compared to MPEG videos. This result in better quality of service for mp4 traffic compared to mpeg traffic. The bandwidth allocation is performed using the following criteria:

$$BWAllot_{mp4} = \begin{cases} BWReq_{mp4} \; if \; BWReq_{mp4} < MRTR, BWAvail \\ MRTR \;\; if \; MRTR < BWReq_{mp4} < \; BWAvail \\ BWAvail \; if \; BWAvail < BWReq_{mp4} < MRTR \end{cases} (1)$$

Where MRTR is the minimum reserve traffic rate.

After an mp4 video is allocated bandwidth, the leftover bandwidth is calculated as:

$$BWAvail = \; BWAvail - \sum_{i=1}^{n} BWAllot_{mp4}^{i} (2)$$

Where n represents the number of mp4 connections that have been allocated bandwidth in the current frame.

Once all the mp4 connections are allotted bandwidth, the connections for MPEG videos are allocated bandwidth as below:

$$BWAllot_{mpeg} = \begin{cases} BWReq_{mpeg} \; if \; BWReq_{mpeg} < MRTR, BWAvail \\ MRTR \;\; if \; MRTR < BWReq_{mpeg} < \; BWAvail \\ BWAvail \; if \; BWAvail < BWReq_{mpeg} < MRTR \end{cases} (3)$$

The available bandwidth is calculated as:

$$BWAvail = \; BWAvail - \sum_{k=1}^{p} BWAllot_{mp4}^{k} (4)$$

Where "p" represents the number of MPEG videos that have already been allocated bandwidth

The term BWAvailinfact represents the available bandwidth after allocating bandwidth to "n" mp4 videos followed by p MPEG videos, i.e.

$$BWAvail = BWAvail - \sum_{i=1}^{n} BWAllot_{mp4}^{i} - \sum_{k=1}^{p} BWAllot_{mp4}^{k} (5)$$

V. Results And Discussion

Simulations were carried out on MATLAB. Simulation parameters are given in the table

TABLE I. SIMULATION PARAMETERS

Parameter	Value
Uplink Bandwidth	10 Mbps
Traffic Type	RTPS high priority data
	nRTPS low priority data
Packet Arrival Pattern	Variable bit rate
Average arrival rate	50 kbps

Simulations were carried out to find the improvement in throughput. Figure 1 shows the throughput results.

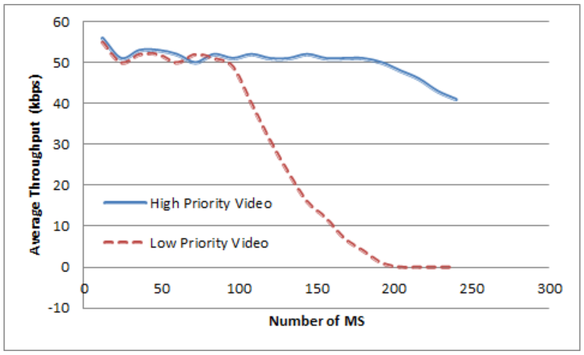


Fig. 1. Average throughput for high priority and low priority video data..

Figure 1 shows that initially there is sufficient bandwidth to support all the MS. Hence both high priority and low priority data show similar throughput results. As the number of MS increases there isn't sufficient bandwidth to support all the traffic since high priority video is classified as RTPS traffic, high priority video gets allocated bandwidth. The left over bandwidth is allocated to the low priority video which is categorized as nRTPS traffic.

As the packets keep accumulating, MS has to decide whether to transmit the delayed data or to drop the packets. In case of video packets delayed transmission results in jittered video which is not suitable for viewing. Hence all delayed packets shall be dropped. Figure 2 shows the simulation results for the data drop in case of high priority and low priority video traffic.

Increase in the number of MS results in more MS contending for limited bandwidth. Since high priority video is classified as RTPS traffic and allotted bandwidth before low priority video, high priority data experiences lower data drop compared to low priority video.

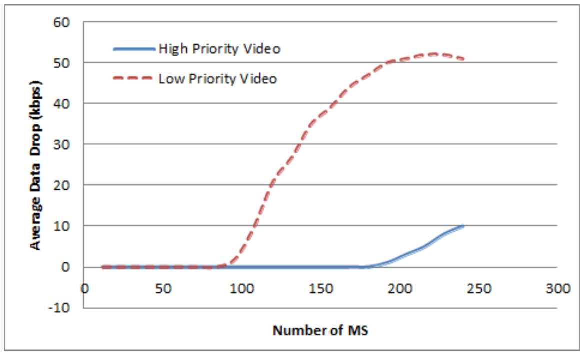


Fig. 2. Average data drop for high priority and low priority video data.

Simulations were carried out to observe the impact of priority of the user and the priority of the data traffic. The order of classification can be as given in Table II.

Data Type	Classification Type
High priority user generating high priority video (HUHT)	RTPS
Low priority user generating high priority video (LUHT)	eRTPS
High priority user generating low priority video (HULT)	nRTPS
Low priority user generating low priority video (LULT)	BE

Figure 3 shows the simulation results for average throughput.

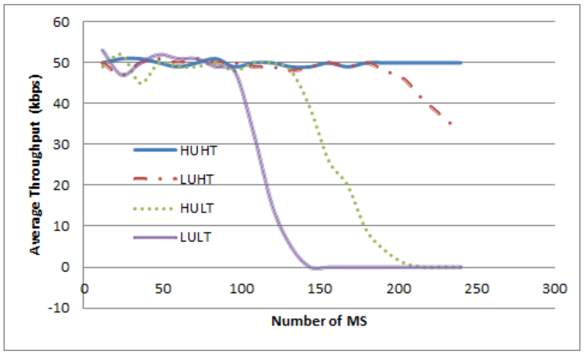


Fig. 3. Average throughput for different classification methods v/s number of MS.

Since HUHT traffic is classified as RTPS traffic, HUHT is allocated bandwidth first then followed by LUHT which is followed by HULT and LULT. As the number of MS increases there isn't bandwidth to support all the video traffic. Hence lower priority traffic is dropped to maintain the throughput of higher priority traffic.

Figure 4 shows the corresponding data drop for the four types of classification.

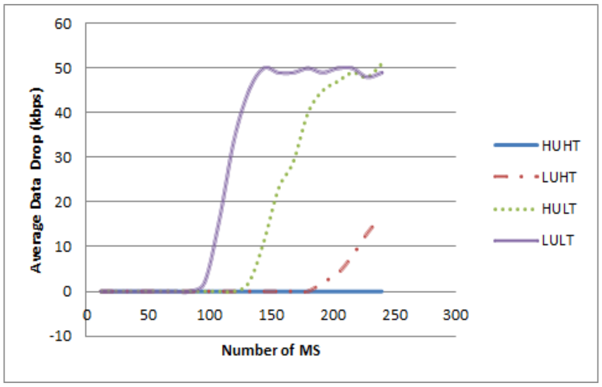


Fig. 4. Average data drop for different classification v/s number of MS.

VI. CONCLUSION

This paper proposes a packet classification algorithm in WiMAX that utilizes the GET and Content type/ MIME type headers of HTTP. By exploiting the application layer headers a fine grained packet classification mechanism can be achieved which in-turn leads to a controlled and coordinated quality of service. Additionally the algorithm is extended to add user based packet classification which provides the necessary flexibility to the network providers. Simulation results show a graded quality of service for video traffic.

REFERENCES

[1] Air Interface for Fixed Broadband Wireless Access Systems, IEEE Std. 802.16, 2004.

[2] Air Interface for Mobile Broadband Wireless Access Systems, IEEE Std. 802.16e, 2005.

[3] IEEE Std. 802.16-2009, IEEE standard for local and metropolitan area networks part 16: air interface for broadband wireless access systems.

[4] RFC 791 http://www.ietf.org/rfc/rfc791.txt.

[5] P. Gupta, N. McKeown, Packet classification on multiple fields, Proceedings of the conference on Applications, technologies, architectures, and protocols for computer communication, pp. 147-160, 1999.

[6] Matlab http://www.mathworks.in/products/matlab/

Design and Development of AlgoWBIs

Kavita
Research Scholar
Banasthali University
Rajasthan, India

Dr. Abdul Wahid
Head of Department of CS & IT
MANUU, Central University
Hyderabad, India

Dr. G N Purohit
Dean, Department of Apaji Institute of Mathematics & Applied Computer Technology, Banasthali University
Rajasthan, India

***Abstract*—AlgoWBIs has been developed to support algorithm learning. The goal to develop this tool is to empower educators and learners with an interactive learning tool to improving algorithm's skills.**

The paper focuses on how to trim down the challenges ofalgorithm learning andalso discusses how this tool will improve the effectiveness of computer facilitated interactive learning and will support in reducing stress of learning. It will aid computer science graduate and undergraduate learners.Perspective development model has been used for development to enhance the tool's features.

***Keywords*—Algorithm, Instructional design; Web-Based Instruction system (WBIs); Meta learning; Self-paced learning; Web-Based learning; Interactive learning; computer based learning; personalized learning**

I. Introduction

Interactive learning system, in contrast to the traditional learning system is more powerful and in favors of entire education system. Interactive learning tools support and allow learners and educators to actively participate in the learning process.

Various interactive learning tools have been developed and are being developed to empower educators in support to change the way teaching and learning occurs[1].Algorithm learning is critical concern for most of the computer science learner and educators. The tool AlgoWBIs is developed to support the instructors and the learners of the information technology and computer science graduate and undergraduate courses. The tool allows students to explore the algorithms for various sorting techniques. The tool not just support learning rather presents simulation, and self assessment approach to test knowledge level[2].

In this paper literature review discusses about the research done this area and summarize with various interactive tools. These tools are available for Electrical Engineering and Electronics, interactive learning modules for PID control, design and development of a web-based interactive software tool for teaching operating systems, DsCats: Animating Data

Structures and PyAlgo as a learning platform for algorithm and data structure.

Further the paper talks about the motivation for this research and descried how and why it is constructive for learners and educators. Model formulation in the subsequent section is discussed. A model which is followed to develop the tool is depicted and described in detail. This section is all about the development process of the tool.

Overall development and the content present in the tool discussed in the consecutive section. This section describes the development of the tool. How the learner can navigate from one topic and page to another one. Finally the paper discusses the limitations and future scope of the tool.

II. Literature review

Justin Cappos & Patrick Homer (2002) Justin Cappos and Patrick Homer has developed Data Structure Computer Animation Tools called DsCats is available for classroom use. This tool supports educator presentations, student experimentation, and programming assignments. It is an animated tool designed for educational use. The tool implements binary search tree, AVL tree, and B-Tree data structures. DsCats implements a number of new features to facilitate its use as a learning tool. The features presented in the tool are features not commonly present in other tools. These include the ability vary the level of detail during the animation, move backward and forward at will through an animation, and the tool display large data sets too.

U. Antonovičs & Ē. Priednieks (2006) reviewed the literature and presented an Interactive Learning Tools for Electrical Engineering and Electronics. They have implemented original student centered interactive learning tools for higher and vocational schools in Latvia. The tool presents laboratory practice simulations, circuit calculation programs, and learning exercises. Here the interactive learning tool is basically designed to support distance learning courses.

The computer-based learning tool has various advantages for the learners. These are interactive learning in the "computer-student" mode, study material is user-friendly, and learners have the opportunity to choose an individual time of learning. Another advantage of the tool is the evaluating and self assessment could be done by the learners and further the education cost is also reduced. Lejla Abazi-Bexheti et al. (2007) reviewed and concluded with an Interactive Multimedia Learning Systems (IMLS). The main objective of the researchers is to create an additional learning tool that combines on-screen text, graphics, animations, audio and video. The researcher concluded with the tool in order to improve the learning process.

Jose Luis Guzman et al. (2008) describes a collection of interactive learning modules for PID control. It is based on the graphical spread sheet metaphor. The modules are designed to speed-up learning and to enhance understanding of the behavior of loops with PID controllers. The modules are implemented in Sysquake, a Matlab dialect with the strong support of interaction. Radio Dragusin and Paula Petcu (2010) designed and develop a tool named PyAlgo for the students following the bachelor level course on Algorithm and Data Structure. The tool focuses on functionality including a library of algorithms and a benchmarking tool.

Aristogiannis Garmpis (2011) after reviewing the literature, researcher developed Design and Development of a Web-based Interactive Software Tool for Teaching Operating Systems. The tool is a complement the existing teaching and learning methods of Operating Systems. The aim of the researcher is not to obsolete the existing pedagogical approach but rather will support the instructor and students for a better understanding of the memory management operations and especially the page replacement algorithm's operation to be used in everyday OS classrooms.The software is intended to support undergraduate students so that they can easily explore the operations of the algorithms through an interaction with the tool. According to the researcher students can explore each algorithm’s mechanism separately using this tool. The students will be able to learn from their mistakes as shown automatically by the software. Sebastian Dormido et al. (2012) Researchers reviewed literature and presented an interactive software tool for the loop shaping design of fractional-order PID controllers. The presented tool allows determining automatically the controller parameters by mapping a point of the process Nyquist plot to a point of the loop transfer function Nyquist plot. According to the researchers this kind of Computer Aided Control System Design tools are very useful from an educational viewpoint and in allowing a widespread use of fractional PID controllers in the industry.

III. Motivation

Bearing lot many advantages over traditional learning interactive learning becoming popular among universities, school and colleges, instructors and the students. The learning process that encourages the learners to think, write and to talk about is known to be an interactive learning. Interactive learning helps the learner to be more dynamic and active.

Keeping interactive learning’s increased requirement in mind literature review is done for interactive algorithm learning tools. Literature supports available tools are learning algorithm. Some are providing learning along with self assessment, some are covering learning and simulation, and few are having only a simulation.Malmi [3] proposed a framework TRAKLA2, for building interactive algorithm simulation exercises only. PyAlgo supports learning platform for algorithm and data structure. The main features of PyAlgo are developing, organizing, and benchmarking algorithms.

An interactive tool named AlgoWBIs for algorithm learning is presented, covering learning, example, simulation, and self assessment all together. Graphical representation of existing tools along with the disclosure of our tool is presented in figure-1. The tool presented here encourages the learners of algorithms to be more active and attentive during learning.

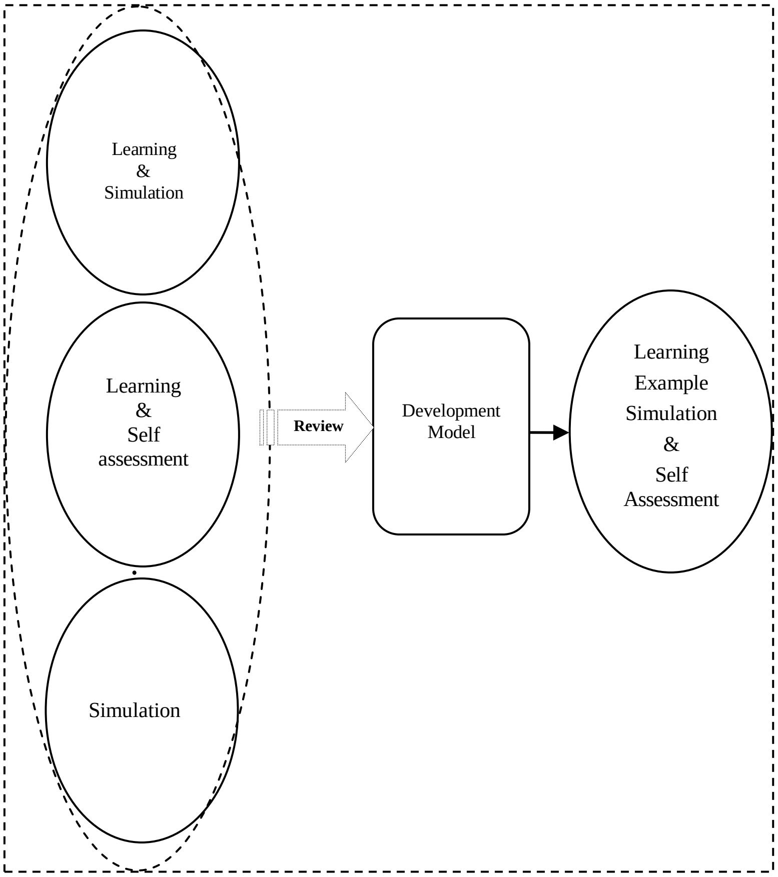


Fig. 1. Review and Development Model

In this figure left side oval shows the available tool for algorithm learning. As shown some of the available tools support learning along with the simulation, learning and self assessment and few supports only simulations. The arrow and rectangle shows that the review is done of the available tools and then follows the development model.

The development model is described in detail in the consecutive section. Finally a new tool for algorithm learning is presented in the right side oval. The new tool is also described in the AlgoWBIs development and content sections.

IV. Model Formulation

One major approach to improve the interactivity of the software is the use of appropriate model. These models are helpful in providing feedback and also supportive in improving and innovative insights for designing and developing excellent interactive system [4]. A model provides information about any activity. That activity Anita Lee-Post e-learning success perspective model, that address the questions of how to design, development, and delivery of successful e-learning systems [5] may include designing, development, installation or testing of any system [6].

After judging, suitability of Anita Lee-Post's perspective model and incremental model and following the model is used (figure-2) is followed to develop this tool.

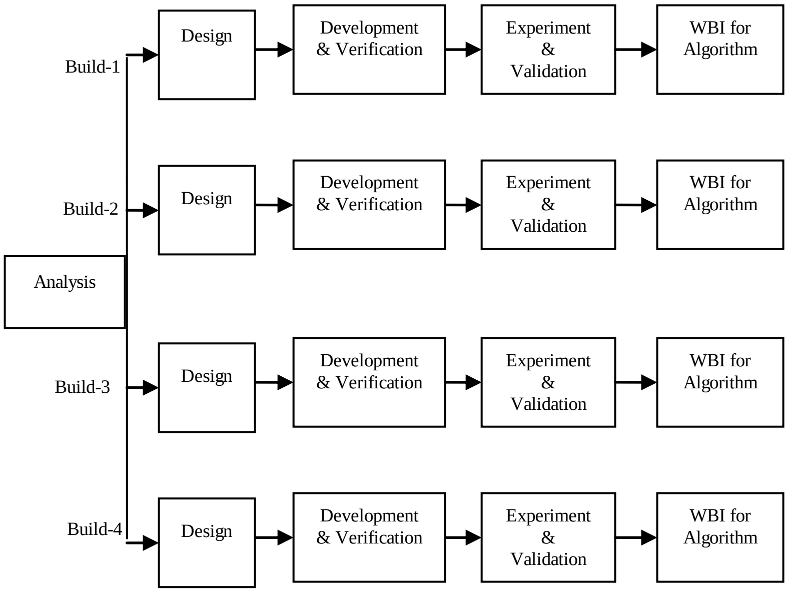


Fig. 2. Perspective Incremental Life cycle model for WBI for Algorithm

Development model used here is based on perspective model, which often can be used as a specification of something to be created. In this model analysis phase allowed us to define the goals and functions to the tool in favor of learners and instructors. Each built sums up with the running software covering the designing, development and verification and experiment and validation. The real code is written in development phase. White box testing is also done in this phase to verify code. Experiment and validation are covering black box testing for each build.

V. AlgoWBIs Development And Content And Features

The major challenge in facilitating educators is providing them a meaningful and interactive learning tool [7]. Development of interactive tool requires diversified knowledge including knowledge of interactive tools along with the programming skills, and command on the subject for which tool is being developed. Knowing the learner's psychology is extremely helpful for the success of such tools [8].

Tools interactivity helps in engaging learners during learning that is why the focus should be on how to increase the interactivity? The main features of interactive tool include interactivity, self assessment, navigation between topics, relevance and coverage of content.

Content is required to be relevant to fulfill the need of the course. The tool AlgoWBIs covers all the relevant content for algorithm learning. It is also considered the depth of the topic is sufficient. "Learning might suffer if the sequence is chosen improperly" [9], For the purpose of improvement in learning it is required to develop a learning tool having the proper sequence of topic coverage, which has been taken care of.

Another focus is given on content's navigation. Learners should be equipped with the freedom to navigate to the topics to some extend [9]. To assess learner's knowledge appropriate quizzes is required with online evaluation. In order to enhance learning Evaluation is done by the tool immediately after learner go through the self assessment. This evaluation supports learners to judge their knowledge level after learning any topic.

Extendibility of the tool is also possible when required. This feature will reduce the cost of learning and development of new algorithm learning tools. The tool addresses the personalized needs of the learners which would nurture their knowledge and skills. Further the system would bridge the gap between the learners and the instructors during learning algorithms and would promote the platform for sharing their ideas and knowledge.

This paper has comprehensive descriptions related to the development of the presented tool. The tool presents the learner with a sequence of choices, each containing sub-choices or sub-menu. The learner can respond by choosing one of the choices given and system performs a corresponding action. Learning options provided with the tool are:

1) *Selection sort*
2) *Bubble sort*
3) *Insertion sort*
4) *Quick sort*

The development of the tool starts with the splash screen. After the splash screen learner's interface presents four learning options as mentioned above. Depending on the current knowledge level individual learner may start learning. Unlike classroom learning each learner independently can start learning rather than forcing them to go with the topic being taught in the classroom.

The tool is developed in such a manner that the learners first will go through the learning of the topic he/she wants to learn. The detailed content is presented to the learners. This includes the basic concept of the topics for the example if learner goes through the learning of the quick sort will first get to know about what a quick sort is?

After learning about a topic, an example is presented for better understanding, during example's presentation if learner wants to go through the topic again s/he can navigate to that. An exercise will be given after going through the example. If the learner is competent enough to complete the given exercise must go through the topic again.

Finally the self assessment quiz is presented which helps the learner to assess their knowledge level. Once the learner clears the one level can move to the further topic. It is clearly defined what action is required to take further. A learner could navigate through a topic is being learned and could navigate to the main menu.

Other than four learning options, menu also has an exit option to quit from the main menu. Each sorting has four sections:

1) ***About sorting*** –This section covers all the aspects of particular sorting. It includes an introduction, algorithm for sorting, sorting function written in C programming language and complexity of sorting. The learner can navigate backward and forward within the given options using navigation buttons. If the learner wants s/he can exit from the topic any time and come to the main menu.

2) ***Example –***Purpose of this section is to investigate the properties of various algorithms and to enhance the knowledge of learner by showing simulation. It allows the students to understand how sorting algorithm works.

3) ***Exercise*** – This section is consisting of a problem provided to learners. To do the exercise learners are required to use knowledge extracted from the prior sections. It is a highly interactive section, allowing learners to make use of knowledge to explore and judge their understanding of the topic. Here learners are asked to move array values to the correct positions interactively. Exercise is the most important option for the learners associated with each sorting.

4) ***Quiz*** – To allow learners to measure their knowledge, tool provides quizzes. This self assessment section allows learners to test knowledge they gather from the previous sections. Each quiz has fifteen questions with the four optional answers. The learner will be asked to select a single answer for each question. Answers of the questions will be evaluated automatically by system and finally result will be displayed to the learner. The result of the quiz will be helpful to let the learner know about the knowledge level [10]. Other than the above mentioned options, an exit option in each topic allows learners to quit from the topic being learned. There is a menu button to allow learners to go on the main menu of the tool. Going through all above mentioned options is very simple as tool is very interactive and user friendly.

The overall functionality of tool is depicted in figure-3. Initially when learner interacts with the system will go through the login verification. After successful verification learning options will appear, which includes insertion sort, quick sort, bubble and selection sort. The figure also has a proposed algorithm to be implemented in future.

All these sorting algorithms have multilayer process in which first layer covers the details about sorting, second layer covers the example for each algorithm whereas third layer, simulation is the most interactive in nature and asks the learners to move array values to the desired positions according to the sorting technique and fourth and last layer allow learners to check their progress in terms of quizzes for each algorithm learned before.

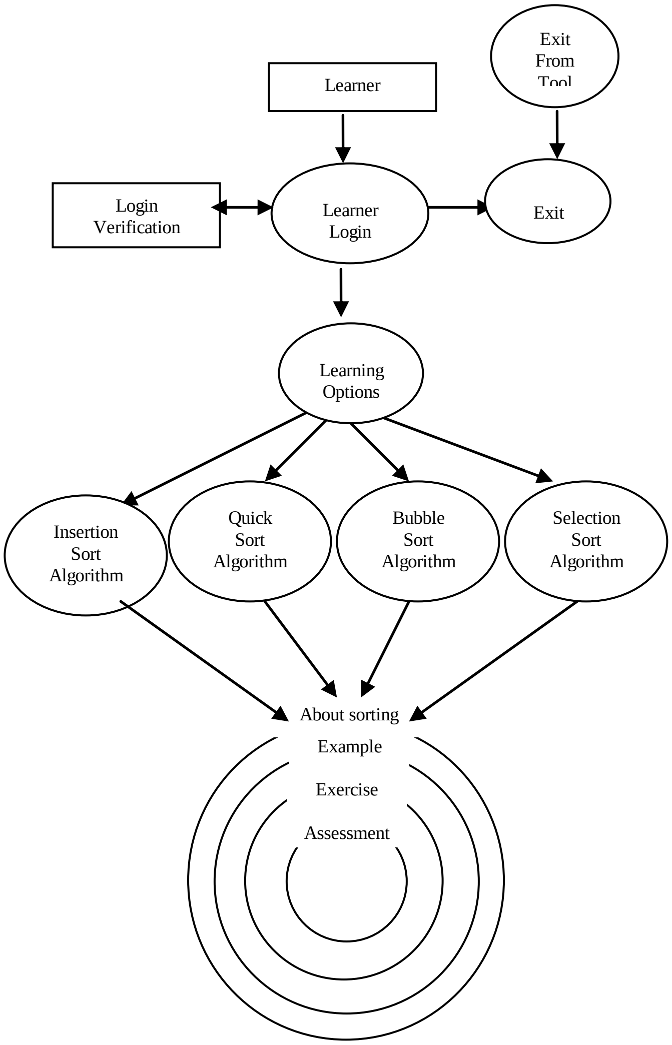


Fig. 3. Functionality of tool AlgoWBIs

VI. Limitations

The AlgoWBIs is developed to support interactive learning with keeping the thing in mind that there may not be full utilization of the tool by the learners and the instructors without proper training, however messages are being prompted whenever required.

Another limitation includes the number of algorithm implementation in the system. Initially the system facilitates the learning for four sorting. Sorting included in are insertion sort, quick sort, bubble and selection sort.

VII. Conclusion and Future work

The study is summarized with a highly interactive tool to overcome the stress of learning and make the learning easy and interactive. The emphasis in this study has been put on the learning of sorting. Modularity and expandability is also kept in mind during development. The tool is developed with Macromedia Authorware, which is deliverable through Network, CD and DVD, Internet etc.

Although this tool fulfills most of the learning requirements, there is always room for improvement. The tool could be expanded in future to cover more algorithms. Further version will cover learning of searching algorithms.

References

[1] Sessoms, D. “Interactive instruction: Creating interactive learning environments through tomorrow’s teachers” International Journal of Technology in Teaching and Learning, Vol, 4(2), pp. 86-96, 2008.

[2] Garmpis, A. “Design and Development of a Web-based Interactive Software Tool for Teaching Operating Systems” Journal of Information Technology Education, Vol. 10, 2011.

[3] Malmi, L., Karavirta, V., Korhonen, A., Nikander, J., Seppälä, O., and Silvasti, P. “Visual Algorithm Simulation Exercise System with Automatic Assessment: TRAKLA2.” Informatics in Education, Vol. 3(2), pp. 267−288,2004.

[4] Arthur, J. D. “A Descriptive/Prescriptive Model for Menu-Based Interaction.” International Journal of Man-Machine Studies, Vol.25(1), pp. 19-32, 1986

[5] Lee-Post, A.(2009). “e-Learning Success Model: an Information Systems Perspective.” Electronic Journal of e-Learning. [Online]. 7(1), pp. 61 – 70. Available: www.ejel.org

[6] Ludewig, J. (2003). “Models in software engineering – an introduction, Softw Syst Model” Vol. 2, pp. 5–14 / Digital Object Identifier (DOI) 10.1007/s10270-003-0020-3

[7] Thomas, R. “Interactivity & Simulationsin e-Learning.” Avaiable: http://www.multiverse.co.uk/whitepaper.pdf

[8] Kennedy, D.M. “SOFTWARE DEVELOPMENT TEAMS IN HIGHER EDUCATION: AN EDUCATOR’S VIEW.”, Software Development Teams in Higher Education: An Educator’s View, 1998.

[9] Abazi-Bexheti, L., Dika, Z. and Luma, A. “Interactive Multimedia Learning Systems: An Experience in Developing IMLS for the IT-Skills Course.” Available: http://rootsitservices.com/CustomPages/sdlifecycle.aspx

[10] Millard,D., Jennings, W., Sanderson A., Wong,A. Patel, A. Brubaker, W., Perala, M. & Slattery, D. “Interactive Learning Modules for Electrical, Computer and Systems Engineering”, 1997.

[11] Nikolaou, A., Koutsouba, M. (2012). “Incorporating 4MAT Model in Distance Instructional Material- An Innovative Design.” European Journal of Open, Distance and E-Learning, Available: http://www.eurodl.org/?article=497

[12] Bellotti, F., Berta, R., De Gloria, A., & Primavera, L. “Supporting authors in the development of task-based learning in serious virtual worlds.” British Journal of Educational Technology, Vol. 41(1), pp. 86–107, 2010.

[13] Hadjerrouit, S. “A Conceptual Framework for Using and Evaluating Web-Based Learning Resources in School Education.” Journal of Information Technology Education, Vol. 9, pp. 53-79, 2010.

[14] Kelly, P., & Stevens, C. (2009). “Narrowing the distance: using e-learner support to enhance the student experience.” European Journal of Open, Distance and E-Learning, Vol. 2. Available: http://www.eurodl.org/materials/contrib/2009/Kelly_Stevens.pdf

[15] Hansen, S.R. & Narayanan, N. H. “On the role of animated analogies in algorithm visualizations.” In: Proceedings of the Fourth International Conference of The Learning Sciences, Lawrence Erlbaum Associates, Mahwah, NJ, pp. 205-211, 2010.

[16] Seay, F. & Catrambone, R. “Using animations to help students learn computer algorithms: A task analysis approach.” Artificial Intelligence in Education, pp. 43-54, 2001.

[17] Craig Larman, Victor R. Basili (2003). "Iterative and Incremental Development: A Brief History". IEEE Computer (IEEE Computer Society), Vol. 36 (6) pp. 47–56. doi:10.1109/MC.2003.1204375. [01-10-2009].

[18] Lee-Post, A. “e-Learning Success Model: an Information Systems Perspective.” Electronic Journal of e-Learning, Vol. 7(1), pp. 61 - 70, Available: www.ejel.org, 2009

[19] Antonovičs, U., Priednieks, Ē. “Interactive Learning Tools for Electrical Engineering and Electronics Course.” Electronics And Electrical Engineering, 2006

[20] Jose´ Luis Guzma´n , Karl J. A stro¨m , Sebastia´n Dormido ,Tore H¨agglund, Yves Piguet. ”Interactive learning modules for PID control.”

[21]]Dormido, S., Pisoni, E. & Visioli, A. “Interactive tools for designing fractional-order PID controllers.” International Journal of Innovative Computing, Information and Control, Vol. 8(7), 2012.

Mining Opinion in Online Messages

Norlela Samsudin
Faculty of Computer and Mathematical Science
Universiti Teknologi MARA
Terengganu, Malaysia

Abdul Razak Hamdan
Faculty of Information Science and Technology,
Universiti Kebangsaan Malaysia
Bangi, Malaysia

Mazidah Puteh
Faculty of Computer and Mathematical Science
Universiti Teknologi MARA,
Terengganu, Malaysia

Mohd Zakree Ahmad Nazri
Faculty of Information Science and Technology,
Universiti Kebangsaan Malaysia
Bangi, Malaysia

***Abstract*— The number of messages that can be mined from online entries increases as the number of online application users increases. In Malaysia, online messages are written in mixed languages known as 'Bahasa Rojak'. Therefore, mining opinion using natural language processing activities is difficult. This study introduces a Malay Mixed Text Normalization Approach (MyTNA) and a feature selection technique based on Immune Network System (FS-INS) in the opinion mining process using machine learning approach. The purpose of MyTNA is to normalize noisy texts in online messages. In addition, FS-INS will automatically select relevant features for the opinion mining process. Several experiments involving 1000 positive movies feedback and 1000 negative movies feedback have been conducted. The results show that accuracy values of opinion mining using Naïve Bayes (NB), k-Nearest Neighbor (kNN) and Sequential Minimal Optimization (SMO) increase after the introduction of MyTNA and FS-INS.**

Keywords—Opinion mining; text normalization; feature selection.

I. INTRODUCTION

It was reported on 30th of Jun 2011, 60.7% or 17.7 million Malaysians used Internet. Facebook is the most favored application [1]. Other than that, communication sites such as blogger.com, mudah.com and Twitter were among the top 10 applications that Malaysians used on the Internet [2]. There is a massive amount of information or opinion that can be gathered from these applications. Nevertheless, very few studies had been conducted to mine opinion from messages that are posted online by Malaysians. The following list demonstrates examples of these messages.

- *"oh bestnya, best giler serius. Nak kasik 5 bintang plus2"*
- *Aku bg 4.9 out of 5stars.Yg 0.1 xcukup to sbb aku xfaham.. masa aku tgk aritu pon xfull"*
- *Ksian ngan kawan aku. Coz abihkan duit utk film nih"*

The examples indicate the following characteristics:

- The use of Malay and English words with Malay words as the main contributors. This scenario is known as Bahasa *Rojak*.
- There is a high number of abbreviations such as *sbb*, *bg* and *tgk*.
- The sentences do not follow the correct syntax of sentence development.

The above scenario make it difficult to mine opinion using natural language processing as expressed by [3]

> "*One drawback of an NLP based approach is that it would likely perform very poorly when used on grammatically incorrect text… methods to detect and possibly correct bad English would be necessary before use on a large scale*."

Furthermore, recognizing subjective words that are relevant to opinion is also a problem in mining opinion using the machine learning approach. The current feature selection techniques in machine learning approach such as Document Frequency (DF), Chi Square and Information Gain assign a value to each feature based on a particular statistical equation.

The features are then sorted. It is up to the user to select the appropriate features based on the sorted value. Different users may select different features. Often, a newbie who is not aware of this scenario would do nothing and causes the classifier transaction to take a longer processing time and use more resources.

The objective of this paper is to introduce a method to normalize noisy texts in Mixed Malay Language texts with the introduction of Malay Mixed Text Normalization Approach (MyTNA). In addition, a new feature selection method named Feature Selection based on Immune Network System (FS-INS) is introduced to select relevant features in opinion mining.

The remainder of the paper is organized as follows: In Section II, previous works in opinion mining using machine learning approach, normalization of noisy text and feature selections in opinion mining process are reviewed. The MyTNA steps and FS-INS algorithm are clarified in Section III. The performance of FS-INS is discussed in Chapter IV. Lastly, conclusion of the study and future research direction are explained in Section V.

II. BACKGROUND

A. Opinion Mining using Machine Learning Approach

Opinions, beliefs, emotions and sentiments are part of private states that cannot be observed. These states are expressed in a document using subjective words [4]. Subjective words that identify the private states may be identified using specific dictionary such as WordNet or SentiWordNet. At the beginning of this century, Pang, Lee and Vaithyanathan [5] started using machine learning approach to mine opinion. Prior to that, opinion mining activities were carried out using natural language processing (NLP) approaches ([6] [7] [8]). Pang, Lee and Vaithyanathan [5] successfully used text mining activities in mining opinion from 700 positive and 700 negative movie reviews. They concluded that additional activities to identify sentiment were required in opinion mining using the machine learning approach. Several researchers used NLP activities in pre-processing steps to select features that are relevant to sentiments ([9] [10] [11]). Other than that, Pang and Lee [12] utilized statistical technique to identify a sentiment phrase. Sentences without sentiment phrases were removed before the opinion mining process. Similarly, Barbosa and Feng [13] used items such as icons and the existence of sentiment words to identify sentiment phrases. Clearly, additional activities in addition to the normal text mining processes are required in opinion mining process. Even though the number of research activities on opinion mining has increased for the past century, none of them studies the performance of opinion mining in Malay language or in bahasa rojak. That is the objective of this paper.

B. Previous works in normalization of noisy texts

Knoblock, Lopresti, Roy and Subramaniam [14] define noisy texts as “any kind of difference between the surface form of a coded representation of the text and the intended, correct or original text”. Before the year 2000, most works on normalization of noisy text involved documents that were created using OMR ([15] [16]). Normalization of noisy texts in short message service (SMS) started to appear in 2005 ([17] [18]). Lately, the normalization of noisy texts has started to use data from online applications such as Twitter messages and Facebook entries. ([19] [20] [21]).

In general, there are three ways to execute the normalization process i.e correction of spelling, machine translation and automatic identification of phonetic. This study uses the first method which includes identifying a noisy text, finding the candidate of correct terms and selecting the correct term.

C. Previous works in feature selection of opinion mining

Feature selection is the process of selecting a set of attributes or features that is relevant to the mining processes. In relation to text mining or opinion mining, every distinct word that exists in the corpus is considered as a feature. The traditional method of feature selection is by selecting all features in a method known as bag of words (BOW). Unfortunately, this method causes certain classifier to perform poorly due to high requirement of resources and longer execution time. Therefore selecting relevant features without reducing the performance of opinion mining process is important. Previous researches in opinion mining use two approaches of feature selection. The first approach uses NLP processes such as Part of Speech (POS) in identifying certain sentence structure or stemming and lemmatization transaction to reduce related forms of a word to a common base form ([5] [9] [22] [23]). The second approach assigns a specific value to every features based on certain statistical equation. Document Frequency uses frequency of the words that exist in the corpus. Information Gain and Mutual Ratio use probability of a word occurring in each class and Chi Square calculates the degree a word is not relevant to a particular class. Unfortunately, these statistical techniques assign a value and sort the features based on these values. It is up to the users to indicate which features should be selected. This study introduces a new feature selection technique that will calculate and select the feature automatically.

III. METHODOLOGY

Fig. 1 illustrates the opinion mining process with the introduction of MyTNA and FS-INS. Both the training data and test data went through normalization process before the opinion mining process.

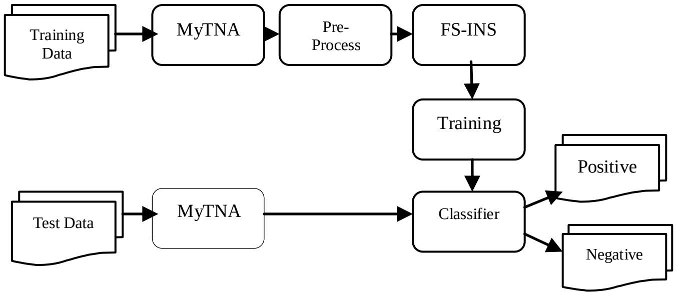


Fig. 1. Opinion mining process for Malay Mixed language

A. Malay Mixed Text Normalization Approach (MyTNA)

The objective of MyTNA is to reduce the number of features by correcting the spellings of common noisy terms and abbreviations that exist in online messages. For example, the word ‘tidak’ is written as ‘tak’, ‘x’, ‘dok’, and ‘dak’ in online messages. When these words are transformed to the features in opinion mining process, there will be five different features that represent the word ‘tidak’. These scenario influences calculation of several classifier such as Naïve Bayes that uses probability calculation in the classification process. Other than that, the value of the word ‘tidak’ in the calculation of k-Nearest Neighbour classifier will be very low since the frequency of the words is 1 instead of 5. Additionally, the word ‘tidak’ is considered as irrelevant if the frequency is low and divided into five different terms instead of one. Therefore, incorrect spellings of words in online messages have to be corrected before the opinion mining process is executed. In this study, the correction of spellings was done using MyTNA

A corpus that consists of 21,000 randomly extracted online messages was derived from e-forum, Facebook and Twitter entries. The following lists were constructed based on this corpus:

This research is supported in part by the Fundamental Research Grant Scheme (FRGS) under the ninth Malaysia Plan (RMK-9), Ministry of Higher Education (MOHE) Malaysia. The grant no is 600-RMI/ST/FRGS 5/3/Fst (208/2010).

- A list of common noisy terms, which consists of noisy terms that exists more than 5 times in the corpus.
- A Bi-Gram list, which consists of the frequency of a word that exists with another word in the corpus.
- A list consists of common English words that are used in the online messages written by the Malaysian. Digital English dictionary is not used in this study because of high similarity between noisy terms in Malay language and the English word such as *'cite'* and *'die'*.

Other than these lists, a list of artificial abbreviations was also derived using the rules that were explained in [24]. During normalization process, each word in the message was tested for out of vocabulary word using a digital Malay dictionary, Common English Word list and Common Acronym list. If the word did not exist in any of these lists, it is considered as a noisy term. The Common Noisy Term list and Artificial Abbreviation list were used to identify potential candidates of the correct word. Later, the Bi-gram list was used to determine the correct translation of the noisy term. MyTNA steps were summarized in Fig. 2.

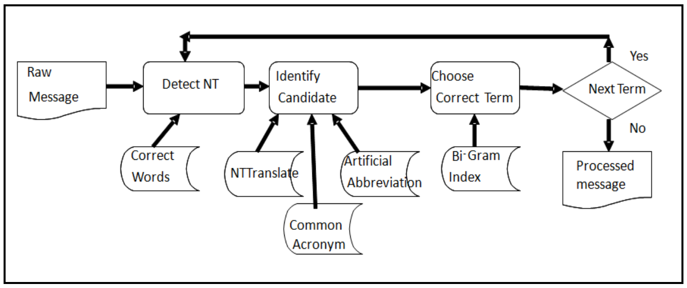


Fig. 2. MyTNA steps

B. Pre-processing

In pre-processing step, features that were not relevant to sentiment were eliminated. The following activities were executed in this study.

- All uppercase letters were changed to small case letters. This is because, online users tend to be creative in writing a message and might write the word 'best' as 'Best', 'BEST', 'BeSt , 'BeST', and 'BesT'. Changing all the uppercase letters into small case letters reduce the number of features.
- All stop words (words without meaning) for Malay language and English language were removed.
- The word 'tidak' was used with the subsequent word. Normally, the word 'tidak' indicates a sentiment. Phrases such as 'tidak best' and 'tidak suka' are synonym with negative sentiments. The word 'tidak' itself exists in positive and negative documents. Normally, the frequency of the word 'tidak' is high in both classes and may be irrelevant to the class. Therefore, using the word 'tidak' with the subsequent word will reduce the frequency of the word 'tidak' itself and increase the number of words that represents the sentiments.

C. Feature Selection based on Immune Network System (FS-INS)

In the filter typed feature selection approach, related features were selected based on certain mathematical equation. Each feature was given a value and later sorted accordingly. However, related features were not selected automatically. Therefore, the users would have to check which features to be selected. If the users were not aware of the activity, all features would be selected, hence resulting poor execution time or the system to crash due to insufficient resources. In term of opinion mining, selecting features that are relevant to positive and negative sentiments is also important. Therefore, in this study, each feature was given a value based on a formula that was introduced by Simeon and Hilderman [25] named Categorical Proportional Difference (CPD). Keefe [26] adjusts the formula to fit the two classes case as shown in Equation 1.

$$CPD(t) = \frac{|FP_i - FN_i\,|}{FP_i + FN_i} \qquad (1)$$

Where

- FP_i is the frequency of term (t) exists in the positive class;
- FN_i is the frequency of term (t) exists in the negative class;

This formula considers a feature as relevant if it occurs in one class in a higher frequency than its frequency in any other classes. For example the feature 'bagus' that exists 100 times in the positive class and 10 times in the negative class will be assigned a value 0.82. On the other hand, a feature that exists in the same frequency in the positive class and negative class is regarded as irrelevant. For example, the term 'saya' that exists 100 times in both classes will be assigned CPD value as 0.0. Therefore, a high CPD value means the feature is very relevant and should be selected. Unfortunately, if a feature exists in only one class, the value 1.0 is assigned to the feature regardless its frequency. Therefore, the relevancy of a feature that exists only once is considered the same as a feature that exists 100 times in only one classAnother problem is to identify the limit of features that is considered as relevance. . In the traditional feature selection techniques, this value is identified by the user based on his or her interpretation of relevancy. To solve these problems, a new feature of selection algorithms based on artificial immune network named FS-INS was created. The main characteristics of this algorithm are listed in the following list:

- If a feature exists in only one class, only the feature that exists more than a certain threshold would be selected.
- A feature is considered as relevant if its CPD's value is above certain threshold.
- A feature is considered relevant if its CPD's value is similar to other features. A feature with CPD's value that matches many other features with similar CPD's value has higher relevancy as compared to other features with less matched of CPD's value.

In the artificial immune system, a feature is considered as a B cell. A memory is used to choose cells with similar CPD's value, while a cell with less matched CPD's value will be deleted from the memory. At the end of the process, all B cells in the memory are considered as the selected features.

D. Experiment's Data

Data for the experiments were randomly collected from various online forums used by Malaysian users such as http://mforum.cari.com.my/portal.php and http://www.mesra.net/forum/. The data were also retrieved from Facebook entries and Twitter messages. Online messages that were selected have the following characteristics.

- It has feedback on a particular movie;
- It has a positive or a negative sentiment; and
- It is written in Malay Mixed Language.

1000 positive movie feedbacks and 1000 negative movie feedbacks were collected and used in all experiments.

E. Experiments

Opinion mining process was executed to analyze the efficiency of MyTNA and FS-INS. k-Nearest Neighbour was used as the classifier in these experiments. Several values of k were tested and it was found that the accuracy value was at the highest in most cases when it is set to 1. Therefore to ensure its consistency, k was set to 1 in all of the experiments. Other than kNN, the accuracy value using Naïve Bayes and Sequential Minimal Optimization were also collected. The accuracy of each opinion mining process was calculated. The accuracy value was calculated using formula in Equation 2

$$Accuracy = \frac{\# \ of \ correct \ prediction}{\# \ of \ reviews} \tag{2}$$

The following experiments were conducted using Weka 3.6 application as the opinion mining tool. Table 1 summarized the experiments for easy understanding.

E1: Opinion mining using raw data;
E2: Opinion mining using raw data and pre-process activities;
E3: Opinion mining using raw data and FS-INS;
E4: Opinion mining using data which had been normalized using MyTNA activities (processed data);
E5: Opinion mining using processed data and pre-process activities;
E6: Opinion mining using processed data, pre-process activities and FS-INS.

TABLE I. LIST OF EXPERIMENTS

Eksp.	Raw Data	MyTNA	Pre Process	FSINS
E1	√			
E2	√		√	
E3	√			√
E4		√		
E5		√	√	
E6		√	√	√

IV. ANALYSIS OF RESULT

The objective of this study is to improve the result of opinion mining messages that are written in 'Bahasa Rojak'. Therefore, the normalisation of noisy text through MyTNA activities and reduction of features using FSINS were applied in the opinion mining process. In addition, several pre-process activities were also conducted prior to the feature selection step. Several experiments were conducted to check the efficiency of these new steps. Table 2 illustrates the result of these experiments.

TABLE II. RESULTS OF EXPERIMENTS

Eksp.	Accuracy		
	NB	*kNN*	*SMO*
E1	85.00	64.10	82.96
E2	87.20	68.00	85.60
E3	86.90	66.70	82.63
E4	85.80	64.07	81.96
E5	89.60	72.60	86.80
E6	91.04	79.08	92.25

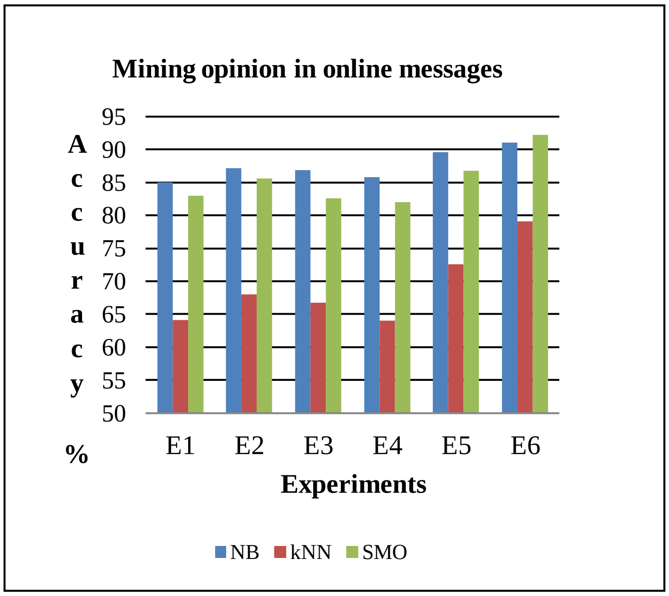


Fig. 3. Results of Experiments

The following conclusions were derived from the results.

- Normalisation of noisy text alone does not improve the opinion mining result. (The result of Experiment E4 is similar to the result of Experiment E1).
- Using only FSINS also does not improve the opinion mining result (The result of Experiment E3 is similar to the result of Experiment E1).
- Combining normalisation of noisy text and the pre-processing activities improves the result of opinion mining slightly (The result of Experiment E5 is better than the result of Experiment E1).
- Combining normalisation of noisy text, the pre-processing activities and using FS-INS as feature selection improves the accuracy value of opinion mining in mixed Malay language (5 % in NB, 15% in kNN and 9% in SMO as shown in Fig. 3.)

It can be concluded that choosing the relevant features improves the result if opinion mining and NB used the probability calculation to predict a class. Additionally, selecting relevant features also improves the probability for predicting the class of new online messages. Similarly, k-Nearest Neighbour classifier uses the class of the nearest neighbour in its prediction. The normalisation of noisy texts process and FS-INS corrects the spelling of most words. In addition, the reduction of features in pre-processing steps and the feature selection technique make it easier for k-Nearest Neighbour to predict the class of a particular message. SMO classifier also creates a virtual line between both classes. Relevant features cause a better line prediction and lead to better accuracy in predicting the appropriate class.

Pang, Lee and Vaithyanathan [5] indicates that additional activities to the normal activities of text mining are required in opinion mining. In this study, several activities were introduced to ensure that only relevant features to sentiments are selected such as

- using of word ‘tidak’ with the subsequent word;
- selection of features based on CPD’s values; and
- selection of features with similar CPD’s values.

These activities contribute to the improvement of accuracy value in opinion mining of online messages written in Malay Mixed Language.

V. CONCLUSION

Improving the accuracy of opinion mining of online messages written in ‘Bahasa Rojak’ is the objective of this study. Executing additional activities such as normalisation of noisy texts approach named MyTNA, several pre-processing activities and a feature selection technique named FS-INS improve the result of opinion mining using NB, kNN and SMO as the classifiers. Nevertheless, more experiments are required to verify whether additional activities introduced in this study improve the opinion mining process. One of them is to validate the result of using FS-INS as feature selection technique as compared to the result of opinion mining using other feature selection techniques such as Document Frequency, Information Gain and Chi Square.

REFERENCES

[1] http://www.internetworldstats.com/stats3.htm, accessed 1/4/2013

[2] http://www.alexa.com/topsites/countries;0/MY, accessed 1/4/2013

[3] O'Neill, A.: ‘Sentiment Mining for Natural Language Documents’, Book Sentiment Mining for Natural Language Documents’, Australian National University, 2009

[4] Wilson, T.W., J.: ‘Annotating Opinions in the World Press’, 2003

[5] Pang, B., Lee, L., and Vaithyanathan, S.: ‘Thumbs up?: sentiment classification using machine learning techniques’. In Proc. of the ACL-02 conference on Empirical methods in natural language processing, 2002, pp. 79-86

[6] Turney, P.D.: ‘Thumbs up or thumbs down?: semantic orientation applied to unsupervised classification of reviews’. In Proc of the 40th Annual Meeting on Association for Computational Linguistics, Philadelphia, Pennsylvania, 2002 pp. 417-428

[7] Nasukawa, T., and Yi, J. ‘Sentiment analysis: capturing favorability using natural language processing’. in Proc. of the 2nd international conference on Knowledge capture, Sanibel Island, FL, USA, 2003 pp. 70-77

[8] http://nlp.stanford.edu/courses/cs224n/2007/fp/johnnyw-hengren.pdf, accessed 21 January 2010 2010

[9] Gamon, M., ‘Sentiment classification on customer feedback data: noisy data, large feature vectors, and the role of linguistic analysis’. In Proc. of the 20th international conference on Computational Linguistics, Geneva, Switzerland, 2004 , pp. 841

[10] Boiy, E., and Moens, P.H. ‘Automatic Sentiment Analysis in On-line Text’, ‘Book Automatic Sentiment Analysis in On-line Text’ 2007, pp. 349-360

[11] Boiy, E., and Moens, M.-F.: ‘A machine learning approach to sentiment analysis in multilingual Web texts’, Information Retrieval, 2009, 12, (5), pp. 526-558

[12] Pang, B., and Lee, L., ‘Seeing stars: exploiting class relationships for sentiment categorization with respect to rating scales’. In Proc. of the 43rd Annual Meeting on Association for Computational Linguistics, Ann Arbor, Michigan, 2005 pp.115-124

[13] Barbosa, L., and Feng, J.: ‘Robust sentiment detection on Twitter from biased and noisy data’. In Proc. of the 23rd International Conference on Computational Linguistics: Posters, Beijing, China 2010 pp. 36-44

[14] Knoblock, C., Lopresti, D., Roy, S., and Subramaniam, L.: ‘Special issue on noisy text analytics’, International Journal on Document Analysis and Recognition, 2007, 10, (3), pp. 127-128

[15] Kernighan, M.D., Church, K.W., and Gale, W.A.: ‘A spelling correction program based on a noisy channel model’. In Proc. of the 13th conference on Computational linguistics - Volume 2, Helsinki, Finland, 1990, pp. 205-210

[16] Mikheev, A.: ‘Document centered approach to text normalization’. In Proc. of the 23rd annual international ACM SIGIR conference on Research and development in information retrieval, Athens, Greece, 2000 pp. 136-143

[17] Aw, A., Zhang, M., Xiao, J., and Su, J., ‘A phrase-based statistical model for SMS text normalization’. In Proc. of the COLING/ACL, Sydney, Australia, 2006 pp. 33-40

[18] Fairon, C.P., S.: ‘A Translated Corpus of 30,000 French SMS’, 2005

[19] Clark, E., and Araki, K.: ‘Text Normalization in Social Media: Progress, Problems and Applications for a Pre-Processing System of Casual English’, Procedia - Social and Behavioral Sciences, 2011, 27, (0), pp. 2-11

[20] Contractor, D., Kothari, G., Faruquie, T.A., Subramaniam, L.V., and Negi, S.: ‘Handling noisy queries in cross language FAQ retrieval’. In Proc. of the 2010 Conference on Empirical Methods in Natural Language Processing, Cambridge, Massachusetts, 2010 pp. 87-96

[21] Laboreiro, G., Lu, Sarmento, s., Teixeira, J., Eug, and Oliveira, n.: ‘Tokenizing micro-blogging messages using a text classification approach’. In Proc. of the fourth workshop on Analytics for noisy unstructured text data, Toronto, Canada, 2010 pp. 81-88

[22] Dave, K., Lawrence, S., and Pennock, D.M.: ‘Mining the peanut gallery: opinion extraction and semantic classification of product reviews’. In Proc. of the 12th international conference on World Wide Web, Budapest, Hungary, 2003 pp. 519-528

[23] Salvetti, F., Reichenbach, C., and Lewis, S.: ‘Opinion Polarity Identification of Movie Reviews Computing Attitude and Affect', in Text: Theory and Applications’, in Shanahan, J., Qu, Y., and Wiebe, J. (Eds.) (Springer Netherlands, 2006), pp. 303-316

[24] [Samsudin, N., Puteh, M., Hamdan, A.R., and Ahmad, M.Z.: ‘Normalization of Common NoisyTerms in Malaysian Online Media’,’In Knowledge Management International Conference (KMICe) , Johor Baharu, 2012, pp. 526 - 531

[25] Simeon, M., and Hilderman, R. ‘Categorical Proportional Difference: A Feature Selection Method for Text Categorization’, in Proc. of Conferences in Research and Practice in Information Technology (CRPIT), 2008, pp. 201-208.

[26] Keefe, T.O.K., I. ‘Feature Selection and Weighting Methods in Sentiment Analysis’, in Proceedings of the 14th Australasian Document Computing Symposium, 2009, pp. 67-74.

English to Creole and Creole to English Rule Based Machine Translation System

Sameerchand Pudaruth, Lallesh Sookun, Arvind Kumar Ruchpaul
Computer Science and Engineering
University of Mauritius
Mauritius

***Abstract*— Machine translation is the process of translating a text from one language to another, using computer software. A translation system is important to overcome language barriers, and help people communicate between different parts of the world. Most of popular online translation system caters only for the most commonly used languages but no research has been made so far concerning the translation of the Mauritian Creole language. In this paper, we present the first machine translation (MT) system that translates English sentences to Mauritian Creole language and vice-versa. The system uses the rule based machine translation approach to perform translation. It takes as input sentences in the source language either in English or Creole and outputs the translation of the sentences in the target language. The results show that the system can provide translation of acceptable quality. This system can potentially benefit many categories of people, since it allow them to perform their translation quickly and with ease.**

Keywords— rule-based; Mauritian Creole; English

I. Introduction

People around the world have access to millions of documents, articles and websites over the Internet. However these electronic documents are often written in different languages and therefore it is necessary to translate them into the relevant target language.

Traditionally, human translators have assisted people to understand texts in a foreign language. However the translation process is very time consuming and costly, when it is done by a human translator. Therefore there is an increasing need for a translation tool so that communication can take place between different parts of the world. A translation tool would help to overcome language barriers.

In Mauritius, even if English is the official language, the majority of the Mauritian population uses the Creole language as a medium of communication. It is a language which is spoken only in Mauritius and some other small islands in the Indian Ocean. Many Mauritians face difficulties in expressing themselves properly using the English language. At the same time, foreigners and most particularly tourists find it extremely difficult to communicate with the Mauritian people.

This is mainly because many Mauritians are at ease only in their native language, which is Creole. The Mauritian Creole language has recently been introduced in the education system in Mauritius [1] as an optional subject and also as a medium of instruction to facilitate teaching and learning in primary schools. The introduction of Creole at school is due to many factors, the most important of which is the high rate of failure at the Certificate of Primary Education (CPE) exams.

Mauritian Creole is thus becoming more and more important for the Mauritian population as it has already been formalized as a full-fledged language. Its usage in formal situations has increased dramatically over the last five years. There was a time when it was considered rude to do advertising in Creole. However, now things have changed considerably and Creole has become the preferred medium for advertising for all range and types of organisations as well as for the government. Even the Bible is now available in Creole.

In this paper we attempt to solve the identified problems by introducing a machine translation system that will help people translate texts from English to Mauritian Creole and vice-versa. The translation system would greatly assist students in switching to and from Mauritian Creole and English language. It would also be vital to foreigners and tourists as it would enable them understand the meaning of certain words or texts in the Creole language.

The tool developed can also be used in conjunction with other translation tools. Thus, A German can use Google translate to translate German texts into English and the use our tool to convert to Creole and vice-versa. Since the Mauritian economy depends heavily on Tourism and tourists from all over the world come to Mauritius, the tool can be of tremendous value to visitors.

This paper is organized as follows. Section 2 looks at the related work. The section is divided into 4 parts: an introduction to machine translation, its architectures and briefly describes the main paradigms that have been developed. An overview of the Mauritian Creole grammar will then be discussed. In section 3, we present how the translation system has been implemented. Finally, we evaluate the system in Section 4 and conclude the paper in Section 5.

II. Related Works

There are a number of issues that needs to be address for rule-based machine translation systems. For example, Charoenpornsawat et al. [2] address the issue of word-sense disambiguation by using machine learning techniques to automatically extract context information from a training corpus.

Their work improves the translation quality of rule based MT systems using this approach. Oliveira et al. [3] shows that by using a systematic approach to break down the length of the sentences based on patterns, clauses, conjunctions, and punctuation can help to parse, and translate long sentences efficiently. Also, Poornima et al. [4] suggest simplifying complex sentences into simpler sentences in order to improve the translation quality. Their research presented results which showed that this method was able to preserve the meaning of the sentence after translation.

A. Machine translation

Machine translation (MT) refers to the process of translating a text from one language (source language) into another language (target language), using computers. The field of MT draws ideas and techniques mainly from linguistics, computer science, artificial intelligence (AI), translation theory and statistics [5]. In recent years, there has been a great increase in activity in the machine translation field, resulting in better translation quality being produced by MT systems. There are numerous motivations towards developing a computer based machine translator. First of all, this can help people perform translation faster and at a lower cost compared to human translators. There is also the perceived need to translate large amount of data and this can be done in a relatively short space of time using MT. Many different MT approaches have been developed, such as rule-based, statistical-based and example-based approaches. However there is still no MT approach that is able to produce high quality translations for broader domain systems. In fact most of the successful MT systems are for a restricted domain, such as the Canadian METEO system [6].

B. Machine translation architectures

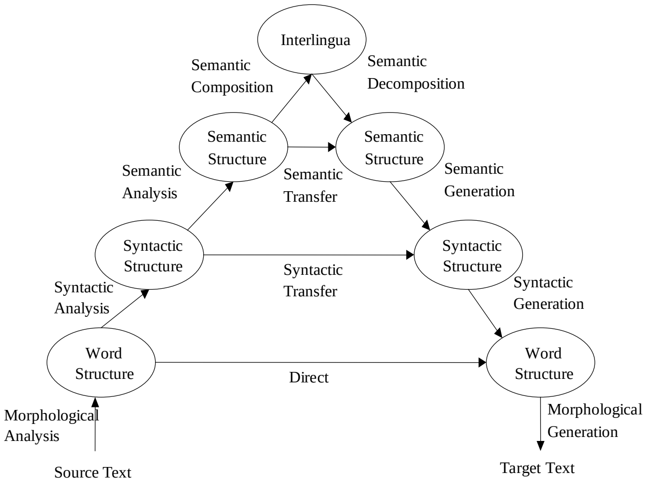


Fig.1. The Vauquois Triangle for MT [8]

Based on the level of linguistic analysis that is processed, MT system may be categorized as: direct, transfer, and Interlingua. The Vauquois triangle as shown in Figure 1 is used to illustrate these three levels. It shows the increasing depth of analysis needed as the top of the triangle is reached (i.e. from direct approach through transfer approach to Interlingua approach) [7]. Furthermore, the amount of transfer knowledge required to traverse the gap between languages decreases as the top of the triangle is reached.

1) Direct Architecture

In direct MT system, morphological analysis is performed on the source text to determine the core of the words to be translated. An example of this is the word "searching" would be analyzed as coming from the word "search". After this stage, each word is translated into a specified language by using large bilingual dictionaries. Then the words are rearranged as according to the target language sentence format.

2) Transfer Architecture

The transfer-based MT architecture was developed to produce better translation quality by capturing and using the linguistic information of the source text. It consists of three stages. During the first stage, the source language (SL) text is analyzed to its syntactic structure (i.e. to produce a parse tree) or semantic structure (i.e. to determine the logical form). Then transfer rules are used to map the SL syntactic/semantic structure to a structure in the target language (TL). Finally in the third stage, the target text is generated from the TL structure [8].

3) Interlingua Architecture

In Interlingua MT system, the source text is analyzed and translated into a language-independent representation known as an Interlingua [9]. The target text is then generated from that Interlingua representation. A common choice of Interlingua used by MT systems is Esperanto.

C. Machine translation paradigms

The main types of MT system are introduced in this section.

1) Rule Based Machine Translation

A Rule-based MT (RBMT) system relies on linguistic rules to perform translation between the source and target language. The rules which are defined in such a system are extensively used in the various processes which analyze an input text, such as during morphological, syntactic and semantic analysis [10].

Rule-based approach is one of the oldest machine translation paradigms that have been developed. RBMT includes concepts such as transfer-based MT and interlingua-based MT. During the translation process in RBMT, the source text is analyzed to produce an intermediate representation (i.e. a parse tree or some abstract representation). The target text is then generated from that intermediate representation [10].

2) Statistical Machine Translation

Statistical MT (SMT) is a MT paradigm in which large bilingual corpora (i.e. a set of translated texts in both the source and target language) are analyzed to produce a translation model, and also large amounts of monolingual text in the target language are used to produce a language model [11]. The MT system then uses the two models to perform translation. The statistical approach has gained a growing interest in recent years, since its re-introduction by a group of IBM researchers in the early nineties. The idea was first proposed by Warren Weaver in 1949, but efforts in this direction were abandoned for various philosophical and theoretical reasons [12].

3) Example-based Machine Translation

Example-Based MT (EBMT) was first suggested by Nagao in 1984 [13]. The basic idea behind EBMT is to generate translation based on previous translation examples. Like statistical MT, example-based MT uses a large bilingual corpus; from which large number of examples are extracted and stored in a database EBMT has three main stages [14]: (i) first of all, the source text is decomposed, and the resulting fragments are matched against a database of real examples, (ii) during the second stage, each fragment is translated and (iii) finally the target text is generated by recombining each fragment.

4) Hybrid Machine Translation

Hybrid approaches to MT integrates various MT paradigms, in an effort to make the most of the strengths of the individual paradigms, while compensating for their weaknesses. The basic idea behind hybrid approaches is to combine linguistic paradigms (i.e. RBMT) with non-linguistic paradigms, such as SMT or EBMT, to produce better results.

D. Mauritian Creole Grammar

The Mauritian Creole (MC) determiner system is quite different from that of French, and it can be considered to be much simpler, as there are no French definite and partitive articles. There is also the exclusion of grammatical gender as well as number in MC.

The core of the MC determiner system has the following functional elements:

- An indefinite singular article **enn** [15].
- A demonstrative **sa**, which is generally used in conjunction with **la** [15].
- A post-nominal specificity marker **la** [15].
- A plural marker **bann** [15].
- The morpheme **li**, which is used to represent the pronoun he/she/it/him/her, depending upon the context it is used.

Mauritian Creole verbs use TMA (Tense, Modality, and Aspect) markers to indicate the tense [16]. The tense marker ‘ti’ indicates an action that has already taken place (i.e. past tense). The modality marker ‘pu’ indicates something will happen (i.e. definite future) whereas the modality marker ‘ava’ is used to express something that may possibly happen (i.e. indefinite future). The aspect marker ‘pe/ape’ marks an action that is still going on (i.e. progressive), in contrast to the aspect marker ‘finn/inn’ which indicates an action that is already over (i.e. perfect) [17].

III. IMPLEMENTATION

The proposed system follows the following steps:

1) Split a text into an array of sentences using “.”, “!”, “?” as delimiters

2) Split each sentence into an array of words using “\W” meta sequence, and the underscore character as delimiters

3) A Greedy algorithm is used to find the longest match for a given fragment, in the database

4) Perform morphological analysis to extract root of word, and check for corresponding translation, in case word has not been translated in step 3.

5) Reorder the words according to the target language sentence format

The rule based machine system relies on the use of bilingual dictionary to perform translation. We have used the Diksioner Morisien dictionary to build the bilingual dictionary in the database.

A. Greedy Algorithm for Natural Language Processing

The greedy algorithm is used to retrieve the target word(s) from the database and it works in following way: it starts at the first character in a sentence, and by traversing from left to right it attempts to find the longest match, based on the words in the database. When a fragment is found, a boundary is marked at the end of the longest match, and then the same searching process continues starting at the next character after the match. If a word is not found, the greedy algorithms remove that character, and then continue the searching process starting at the next character [18]. The steps to perform the greedy search are given below.

```
Initialize startingPos to 0
Initialize numElementsWordArray to number of elements in wordArray
Initialize fragment to 5
Create an array translatedWordArray to Store target word
Create an array wordCategoryArray to Store word category
WHILE starting position <numElementsWordArray
        Initalize flag to 0
        fragment = fragment -1
Initialize fragmentEndPos to startingPos + fragment
IF fragmentEndPos>numElementsWordArray THEN
        fragmentEndPos = numElementsWordArray
ENDIF
Create an empty String greedyString to Store the fragment of words
FOR (k= startingPos; k <fragmentEndPos; increment k)
        greedyString = greedyString + “ ” + wordArray [k]
ENDFOR
IF flag is 0, and target word and word category is retrieved from database where source word = greedyString THEN
        Put target word in translatedWordArray
        Put word category in wordCategoryArray
        Set startingPos to fragmentEndPos
        Set fragment to 5
        Set flag to 1
ENDIF
IF flag is 0, and word not found THEN
        CALL morphologicalAnalysis (greedyString)
        Store the variables received from the morphologicalAnalysis function in a List with parameters (target word, word category)
        Put target word in translatedWordArray
        Put word category in wordCategoryArray
        Set startingPos to fragmentEndPos
```

```
30.         Set fragment to 5
31.     ENDIF
32.ENDWHILE
```

The above greedy algorithm starts the searching process by taking a fragment of maximum of four words starting from the beginning of a sentence. If the fragment is not found in the database, the last word of the fragment is removed, and the searching process continues. If a fragment is found, a boundary is marked at its end, and the searching process continues taking another fragment of a maximum of four words, starting from the boundary.

B. *Morphological Analysis*

In case a word is not translated after the Greedy search sub-process, it enters the Morphological analysis process, where rules are applied to the word to find its root, which is then searched in the database. An example of a morphological rule is: the suffix –ing is removed from the word walking to obtain its root (i.e. walk). The steps for the morphological analysis process are as follows:

```
1.  Get word
2.  Set String truncatedWord to word
3.  Create an empty String saveLastChar to Store the last
    characters of a word
4.  Create an empty String translatedWordString to Store
    translated word
5.  Create an empty String wordCategoryString to Store word
    category
6.  IF length of word > 3 THEN
7.      FOR num= 1to 4
8.          Set truncatedChar to last character of
            truncatedWord
9.          Add truncatedChar to saveLastChar
10.     truncatedWord = truncatedWord – last character
11.     IF num = 1 THEN
12.         IF the reverse of String saveLastChar =
            suffix THEN
13.             IF target word and word
        category is retrieved from database where source
        word = truncatedWord THEN
14.                 Add data to
                translatedWordString (optional)
15.                 Add target word to
                translatedWordString
16.                 Add data to
                translatedWordString (optional)
17.                 Add word category to
                wordCategoryString
18.                 Break
19.             ENDIF
20.         ENDIF
21.     ENDIF
22.     IF num = 2 THEN
23.         Repeat steps 12-20
24.     ENDIF
25.     IF num = 3 THEN
26.         Repeat steps 12-20
27.     ENDIF
28.     IF num = 4 THEN
29.         Repeat steps 12-19
30.         Add the reverse of String saveLastChar
            to truncatedWord
31.         Add the uppercase of String
            truncatedWord to translatedWordString
32.             Add data to wordCategoryString
33.             Break
34.             ENDIF
35.         ENDIF
36.     ENDFOR
37.ELSE
38.     Add the uppercase of String word to translatedWordString
39.     Add data to wordCategoryString
40.ENDIF
41.RETURN the variables (translatedWordString,
wordCategoryString) in an Array
```

C. *Reorder word*

After the words are translated, they are rearranged according to the target language sentence format. The pseudocode for reordering words is given below:

```
1. Get translatedWordArray and wordCategoryArray
2. FOR EACH element in wordCategoryArray
3.      Initialize key to the key of the array element
4.      Initialize value to the value of the array element
5.      IF key is not equal to 0
6.          IF wordCategoryArray [key -1] is equal
            to adjective, and value equal to noun
7.          CALL swapArrayElement
            (translatedWordArray, key-1, key)
8.          CALL swapArrayElement
            (wordCategoryArray, key-1, key)
9.          ENDIF
10.     ENDIF
11.ENDFOR EACH
12.RETURN the variables (translatedWordArray,
wordCategoryArray) in an Array
```

IV. Evaluation

Our goal is to provide the most accurate translation; therefore, whenever new rules were added, a series of tests was carried out to make sure that it does not affect the quality of translation.

Table 1 presents some sample translations obtained when translating sentences from English to Mauritian Creole.

TABLE I. TRANSLATION OF ENGLISH SENTENCES TO MAURITIAN CREOLE

Source text	Expected Result	Target Text
She is a brilliant student	Li enn zelev intelizan	Li ene zelev intelizan
I love spicy food	Mo kontan manze epise	Mo kontan manze epise
I can't tell if he is listening to me or not	Mo pa capav dir si lip ekoute mwa oubien non	Mo pa capav dir si li pe ekoute mwa oubien pa
Either take it or leave it	Swa pran li oubien les li	Swa pran li oubien dekale li

From the above table, we have shown that the translation obtained is of acceptable quality. The column 'Expected result' represents the result obtained from a manual translator while 'Target Text' is the text generated from the program. However for some cases, for e.g. in the last sentence, the word "leave" is being translated to the word **dekale** in Mauritian Creole, which is being used in the wrong context. The Creole word **dekale** means 'put it somewhere else' or 'move it slightly' in English. This will be remedied in our future work where context will be used to deal with polysymy, i.e., words that have different meanings.

Table 2 presents some sample translation obtained when translating from Mauritian Creole to English.

TABLE II. TRANSLATION OF MAURITIAN CREOLE SENTENCES TO ENGLISH

Source text	Expected Result	Target Text
Done to disan, sauve ene lavi!	Give your blood, save a life!	Give your blood, save a life!
Apel mwa kan ou rente lakaz	Call me when you get home	Call me when you enter house
Dan dimans nou mete promotion lor diri	On Sunday we offer discount on rice	On sunday we put promotion on rice
Divin bon pou lasante	Wine is good for health	Wine good for health

As can be seen, most of the generated text is similar to those obtained by the human translator. In the last example, the word 'is' is missing as it is currently quite difficult to know when to add these types of words when translating from English to Creole. The rules are quite complex and there are too many exceptions. It is also challenging to deal with cases of synonymy, i.e. one word in the English language can be translated to several words with completely different meanings in Creole.

A. Weaknesses of the current system

Word sense disambiguation (WSD) is the process of identifying the appropriate sense of a word in a given context. This has not been addressed. Another weakness is that the system can deal with only short sentences. Thus, in our future work, we intend to improve the translation for longer sentences as well. The tool can also be converted into a web application so that it is easily accessible to everyone.

V. CONCLUSION

In this paper, we have implemented the first automated translation system that performs translation of English sentences to Mauritian Creole and vice-versa. The translation system uses the rule-based machine translation approach to perform translation. The results obtained show that the implemented system can provide translation of acceptable quality. The speed of translation of the system is also satisfactory. As part of our future work, we plan to investigate how the problem of word sense disambiguation can be solved and how translation can be improved for longer sentences. The translation system would benefit both foreigners and the Mauritian population as it would enable them to swap between their mother tongue and Mauritian Creole with ease and convenience.

REFERENCES

[1] QUIRIN, S., 2012. Kreol Morisien au Primaire. Week End, 15 Jan. p17.

[2] CHAROENPORNSAWAT, P. AND SORNLERTLAMVANICH, V. AND CHAROENPORN, T., 2002. Improving translation quality of rule-based machine translation. 19th International Conference on Computational Linguistics.

[3] OLIVEIRA, F. AND WONG, F. AND HONG, I., 2010. Systematic Processing of Long Sentences in Rule Based Portuguese-Chinese Machine Translation. 11th Annual Conference on Computational Linguistics and Intelligent Text Processing, 21-27 March 2010, Iasi, Romania. Heidelberg: Springer, pp. 417-426.

[4] POORNIMA, C. AND DHANALAKSHMI, V. AND ANAND, K.M. AND SOMAN, K.P., 2011. Rule based Sentence Simplification for English to Tamil Machine Translation System. International Journal of Computer Applications, 25 (8), 38-42.

[5] HUTCHINS, W.J. AND SOMERS, H.L., 1992. An Introduction to Machine Translation. London: Academic Press.

[6] CHANDIOUX, J., 1976. METEO, an operational system for the translation of public weather forecasts. Seminar on Machine Translation, 8-9 March 1976, Rosslyn, Virginia. Stroudsburg: Association for Computational Linguistics, 27–36.

[7] JURAFSKY, D. AND MARTIN, J.H., 2009. Speech and language processing: an introduction to natural language processing, computational linguistics, and speech recognition. 2nd ed. New Jersey: Prentice Hall.

[8] NIE, J-Y., 2010. Cross-Language Information Retrieval. San Rafael, California: Morgan & Claypool Publishers.

[9] MISHRA, R.B., 2011. Artificial Intelligence PB. New Delhi: PHI Learning Private Limited.

[10] CARL, M. AND WAY, A., 2003. Recent Advances in Example-based Machine Translation. Heidelberg: Springer.

[11] UEFFING, N. AND HAFFARI, G. AND SARKAR, A., 2007. Semi-supervised learning for Machine Translation. Machine Translation, 21 (2), 77-94.

[12] ARMSTRONG, A-W. AND ARMSTRONG, S., 1994. Using large corpora. Cambridge: MIT Press.

[13] NAGAO, M., 1984. A Framework of a Mechanical Translation between Japanese and English by Analogy Principle. Artificial and human intelligence. Elsevier Science Publishers, pp. 173-180.

[14] Somers, H., 2003. An overview of EBMT. In: M. CARL AND A. WAY, ed. Recent Advances in Example-based Machine Translation. Heidelberg: Springer, 3-57.

[15] GUILLEMIN, D., 2011. The Syntax and Semantics of a Determiner System: A Case Study of Mauritian Creole. Amsterdam: John Benjamins Publishing Company.

[16] ADONE, D., 1994. The acquisition of Mauritian Creole. Amsterdam: John Benjamins Publishing Company.

[17] DICK, J-Y. AND AH-VEE, A. AND COLLEN, L., 2011. A Few Points on Verbs in Mauritian Kreol and in Creole Languages [online]. Port-Louis, Ledikasyon pu Travayer.

[18] PALMER, David D., 1997. A trainable rule-based algorithm for word segmentation. Proceedings of the 35th annual meeting on Association for Computational Linguistics, 7-12 July 2007, Madrid, Spain.

A Hybrid Method to Improve Forecasting Accuracy Utilizing Genetic Algorithm –An Application to the Data of Operating equipment and supplies

Daisuke Takeyasu
The Open University of Japan, 2-11 Wakaba,
Mihama-District, Chiba City, 261-8586,
Japan

Kazuhiro Takeyasu†
College of Business Administration, Tokoha
University,325 Oobuchi, Fuji City, Shizuoka, 417-0801,
Japan

***Abstract*—In industries, how to improve forecasting accuracy such as sales, shipping is an important issue. There are many researches made on this. In this paper, a hybrid method is introduced and plural methods are compared. Focusing that the equation of exponential smoothing method(ESM) is equivalent to (1,1) order ARMA model equation, new method of estimation of smoothing constant in exponential smoothing method is proposed before by us which satisfies minimum variance of forecasting error. Generally, smoothing constant is selected arbitrarily. But in this paper, we utilize above stated theoretical solution. Firstly, we make estimation of ARMA model parameter and then estimate smoothing constants. Thus theoretical solution is derived in a simple way and it may be utilized in various fields. Furthermore, combining the trend removing method with this method, we aim to improve forecasting accuracy. An approach to this method is executed in the following method. Trend removing by the combination of linear and 2nd order non-linear function and 3rd order non-linear function is executed to the data of Operating equipment and supplies for three cases (An injection device and a puncture device, A sterilized hypodermic needle and A sterilized syringe). The weights for these functions are set 0.5 for two patterns at first and then varied by 0.01 increment for three patterns and optimal weights are searched. Genetic Algorithm is utilized to search the optimal weight for the weighting parameters of linear and non-linear function. For the comparison, monthly trend is removed after that. Theoretical solution of smoothing constant of ESM is calculated for both of the monthly trend removing data and the non-monthly trend removing data. Then forecasting is executed on these data. The new method shows that it is useful for the time series that has various trend characteristics and has rather strong seasonal trend. The effectiveness of this method should be examined in various cases.**

Keywords—minimum variance; exponential smoothing method; forecasting; trend; operating equipment and supplies

I. INTRODUCTION

Time series analysis is often used in such themes as sales forecasting, stock market price forecasting etc. Sales forecasting is inevitable for Supply Chain Management. But in fact, it is not well utilized in industries. It is because there are so many irregular incidents therefore it becomes hard to make sales forecasting. A mere application of method does not bear good result. The big reason is that sales data or production data are not stationary time series, while linear model requires the time series as a stationary one. In order to improve forecasting accuracy, we have devised trend removal methods as well as searching optimal parameters and obtained good results. We created a new method and applied it to various time series and examined the effectiveness of the method. Applied data are sales data, production data, shipping data, stock market price data, flight passenger data etc.

Many methods for time series analysis have been presented such as Autoregressive model (AR Model), Autoregressive Moving Average Model (ARMA Model) and Exponential Smoothing Method (ESM)[1]−[4]. Among these, ESM is said to be a practical simple method.

For this method, various improving method such as adding compensating item for time lag, coping with the time series with trend[5], utilizing Kalman Filter[6], Bayes Forecasting[7], adaptive ESM[8], exponentially weighted Moving Averages with irregular updating periods [9], making averages of forecasts using plural method [10] are presented. For example, Maeda[6] calculated smoothing constant in relationship with S/N ratio under the assumption that the observation noise was added to the system. But he had to calculate under supposed noise because he could not grasp observation noise. It can be said that it doesn’t pursue optimum solution from the very data themselves which should be derived by those estimation.

Ishii [11] pointed out that the optimal smoothing constant was the solution of infinite order equation, but he didn’t show analytical solution. Based on these facts, we proposed a new method of estimation of smoothing constant in ESM before [12],[20]. Focusing that the equation of ESM is equivalent to (1,1) order ARMA model equation, a new method of estimation of smoothing constant in ESM was derived. Furthermore, combining the trend removal method, forecasting accuracy was improved, where shipping data, stock market price data etc. were examined [13]－[20].

In this paper, utilizing above stated method, a revised forecasting method is proposed. In making forecast such as production data, trend removing method is devised. Trend removing by the combination of linear and *2nd* order non-linear function and *3rd* order non-linear function is executed to the data of Operating equipment and supplies for three cases (An injection device and a puncture device, A sterilized hypodermic needle and A sterilized syringe). These Operating equipment and supplies are used for medical use. The weights for these functions are set 0.5 for two patterns at first and then varied by 0.01 increments for three patterns and optimal weights are searched. Genetic Algorithm is utilized to search the optimal weight for the weighting parameters of linear and non-linear function. For the comparison, monthly trend is removed after that. Theoretical solution of smoothing constant of ESM is calculated for both of the monthly trend removing data and the non monthly trend removing data. Then forecasting is executed on these data. This is a revised forecasting method. Variance of forecasting error of this newly proposed method is assumed to be less than those of previously proposed method. The rest of the paper is organized as follows. In section 2, ESM is stated by ARMA model and estimation method of smoothing constant is derived using ARMA model identification. The combination of linear and non-linear function is introduced for trend removing in section 3. The Monthly Ratio is referred in section 4. Forecasting Accuracy is defined in section 5. Optimal weights are searched in section 6. Forecasting is carried out in section 7, and estimation accuracy is examined.

II. DESCRIPTION OF ESM USING ARMA MODEL [12]

In ESM, forecasting at time $t+1$ is stated in the following equation

$$\hat{x}_{t+1} = \hat{x}_t + \alpha(x_t - \hat{x}_t) \tag{1}$$

$$= \alpha x_t + (1-\alpha)\hat{x}_t \tag{2}$$

Here,

$\hat{x}_{t+1}$: forecasting at $t+1$

x_t : realized value at t

α : smoothing constant $(0 < \alpha < 1)$

(2) is re-stated as

$$\hat{x}_{t+1} = \sum_{l=0}^{\infty} \alpha(1-\alpha)^l x_{t-l} \tag{3}$$

By the way, we consider the following (1,1) order ARMA model.

$$x_t - x_{t-1} = e_t - \beta e_{t-1} \tag{4}$$

Generally, (p,q) order ARMA model is stated as

$$x_t + \sum_{i=1}^{p} a_i x_{t-i} = e_t + \sum_{j=1}^{q} b_j e_{t-j} \tag{5}$$

Here,

MA process in (5) is supposed to satisfy convertibility condition. Utilizing the relation that

$$E[e_t | e_{t-1}, e_{t-2}, \cdots] = 0$$

we get the following equation from (4)

$$\hat{x}_t = x_{t-1} - \beta e_{t-1} \tag{6}$$

Operating this scheme on $t+1$, we finally get

$$\begin{aligned} \hat{x}_{t+1} &= \hat{x}_t + (1-\beta)e_t \\ &= \hat{x}_t + (1-\beta)(x_t - \hat{x}_t) \end{aligned} \tag{7}$$

If we set $1-\beta = \alpha$, the above equation is the same with (1), i.e., equation of ESM is equivalent to (1,1) order ARMA model, or is said to be (0,1,1) order ARIMA model because 1st order AR parameter is -1. Comparing with (4) and (5), we obtain

$$\begin{cases} a_1 = -1 \\ b_1 = -\beta \end{cases}$$

From (1), (7),

$$\alpha = 1 - \beta$$

Therefore, we get

$$\begin{cases} a_1 = -1 \\ b_1 = -\beta = \alpha - 1 \end{cases} \tag{8}$$

From above, we can get estimation of smoothing constant after we identify the parameter of MA part of ARMA model. But, generally MA part of ARMA model become non-linear equations which are described below.

Let (5) be

$$\tilde{x}_t = x_t + \sum_{i=1}^{p} a_i x_{t-i} \tag{9}$$

$$\tilde{x}_t = e_t + \sum_{j=1}^{q} b_j e_{t-j} \tag{10}$$

We express the autocorrelation function of $\tilde{x}_t$ as $\tilde{r}_k$ and from (9), (10), we get the following non-linear equations which are well known.

$\{x_t\}$: Sample process of Stationary Ergodic Gaussian Process $x(t)$ $t = 1,2,\cdots,N,\cdots$

$\{e_t\}$: Gaussian White Noise with 0 mean σ_e^2 variance

$$\left.\begin{array}{l}\begin{cases}\tilde{r}_k = \sigma_e^2 \sum_{j=0}^{q-k} b_j b_{k+j} & (k \le q) \\ 0 & (k \ge q+1)\end{cases} \\ \tilde{r}_0 = \sigma_e^2 \sum_{j=0}^{q} b_j^2\end{array}\right\} \quad (11)$$

For these equations, recursive algorithm has been developed. In this paper, parameter to be estimated is only b_1, so it can be solved in the following way.

From (4) (5) (8) (11), we get

$$\left.\begin{array}{l} q = 1 \\ a_1 = -1 \\ b_1 = -\beta = \alpha - 1 \\ \tilde{r}_0 = \left(1 + b_1^2\right)\sigma_e^2 \\ \tilde{r}_1 = b_1 \sigma_e^2 \end{array}\right\} \quad (12)$$

If we set

$$\rho_k = \frac{\tilde{r}_k}{\tilde{r}_0} \quad (13)$$

the following equation is derived

$$\rho_1 = \frac{b_1}{1 + b_1^2} \quad (14)$$

We can get b_1 as follows

$$b_1 = \frac{1 \pm \sqrt{1 - 4\rho_1^2}}{2\rho_1} \quad (15)$$

In order to have real roots, ρ_1 must satisfy

$$\left|\rho_1\right| \le \frac{1}{2} \quad (16)$$

From invertibility condition, b_1 must satisfy

$$\left|b_1\right| < 1$$

From (14), using the next relation,

$$\left(1 - b_1\right)^2 \ge 0$$
$$\left(1 + b_1\right)^2 \ge 0$$

(16) always holds

As

$$\alpha = b_1 + 1$$

b_1 is within the range of

$$-1 < b_1 < 0$$

Finally we get

$$\left.\begin{array}{l} b_1 = \dfrac{1 - \sqrt{1 - 4\rho_1^2}}{2\rho_1} \\ \alpha = \dfrac{1 + 2\rho_1 - \sqrt{1 - 4\rho_1^2}}{2\rho_1} \end{array}\right\} \quad (17)$$

which satisfies above condition. Thus we can obtain a theoretical solution by a simple way. Focusing on the idea that the equation of ESM is equivalent to (1,1) order ARMA model equation, we can estimate smoothing constant after estimating ARMA model parameter. It can be estimated only by calculating 0th and 1st order autocorrelation function.

III. Trend Removal Method

As trend removal method, we describe the combination of linear and non-linear function.

[1] Linear function

We set

$$y = a_1 x + b_1 \quad (18)$$

as a linear function.

[2] Non-linear function

We set

$$y = a_2 x^2 + b_2 x + c_2 \quad (19)$$

$$y = a_3 x^3 + b_3 x^2 + c_3 x + d_3 \quad (20)$$

as a 2nd and a 3rd order non-linear function. (a_2, b_2, c_2) and (a_3, b_3, c_3, d_3) are also parameters for a 2nd and a 3rd order non-linear functions which are estimated by using least square method.

[3] The combination of linear and non-linear function.

We set

$$\begin{array}{l} y = \alpha_1\left(a_1 x + b_1\right) + \alpha_2\left(a_2 x^2 + b_2 x + c_2\right) \\ \quad + \alpha_3\left(a_3 x^3 + b_3 x^2 + c_3 x + d_3\right) \end{array} \quad (21)$$

$$0 \le \alpha_1 \le 1, 0 \le \alpha_2 \le 1, 0 \le \alpha_3 \le 1, \alpha_1 + \alpha_2 + \alpha_3 = 1 \quad (22)$$

as the combination linear and 2nd order non-linear and 3rd order non-linear function. Trend is removed by dividing the original data by (21). The optimal weighting parameter

$\alpha_1, \alpha_2, \alpha_3$,are determined by utilizing GA. GA method is precisely described in section 6.

IV. Monthly Ratio

For example, if there is the monthly data of L years as stated bellow:

$$\{x_{ij}\}\ (i=1,\cdots,L)\ (j=1,\cdots,12)$$

Where, $x_{ij} \in R$ in which j means month and i means year and x_{ij} is a shipping data of i -th year, j -th month. Then, monthly ratio $\tilde{x}_j\ (j=1,\cdots,12)$ is calculated as follows.

$$\tilde{x}_j = \frac{\frac{1}{L}\sum_{i=1}^{L} x_{ij}}{\frac{1}{L}\cdot\frac{1}{12}\sum_{i=1}^{L}\sum_{j=1}^{12} x_{ij}} \tag{23}$$

Monthly trend is removed by dividing the data by (23). Numerical examples both of monthly trend removal case and non-removal case are discussed in 7.

V. Forecasting Accuracy

Forecasting accuracy is measured by calculating the variance of the forecasting error. Variance of forecasting error is calculated by:

$$\sigma_\varepsilon^2 = \frac{1}{N-1}\sum_{i=1}^{N}(\varepsilon_i - \bar{\varepsilon})^2 \tag{24}$$

Where, forecasting error is expressed as:

$$\varepsilon_i = \hat{x}_i - x_i \tag{25}$$

$$\bar{\varepsilon} = \frac{1}{N}\sum_{i=1}^{N}\varepsilon_i \tag{26}$$

VI. Searching Optimal Weights Utilizing GA

A. Definition of the problem

We search $\alpha_1, \alpha_2, \alpha_3$ of (21) which minimizes (24) by utilizing GA. By (22), we only have to determine α_1 and α_2. $\sigma_\varepsilon{}^2$ ((24)) is a function of α_1 and α_2, therefore we express them as $\sigma_\varepsilon^2(\alpha_1, \alpha_2)$. Now, we pursue the following:

$$\begin{aligned}&\text{Minimize: } \sigma_\varepsilon^2(\alpha_1,\alpha_2)\\ &\quad \text{subject to: } 0 \le \alpha_1 \le 1, 0 \le \alpha_2 \le 1, \alpha_1 + \alpha_2 \le 1\end{aligned} \tag{27}$$

We do not necessarily have to utilize GA for this problem which has small member of variables. Considering the possibility that variables increase when we use logistics curve etc in the near future, we want to ascertain the effectiveness of GA.

B. The structure of the gene

Gene is expressed by the binary system using {0,1} bit. Domain of variable is [0,1] from (22).

We suppose that variables take down to the second decimal place. As the length of domain of variable is 1-0=1, seven bits are required to express variables. The binary bit strings <bit6, ～,bit0> is decoded to the [0,1] domain real number by the following procedure.[21]

Procedure 1: Convert the binary number to the binary-coded decimal.

$$\begin{aligned}&(\langle bit_6, bit_5, bit_4, bit_3, bit_2, bit_1, bit_0\rangle)_2\\ &= \left(\sum_{i=0}^{6} bit_i 2^i\right)_{10}\\ &= X'\end{aligned} \tag{28}$$

Procedure 2: Convert the binary-coded decimal to the real number.

$$\begin{aligned}&\text{The real number}\\ &= \text{(Left hand starting point of the domain)}\\ &+ X' \text{((Right hand ending point of the domain)/(} 2^7 - 1\text{))}\end{aligned} \tag{29}$$

The decimal number, the binary number and the corresponding real number in the case of 7 bits are expressed in Table 6-1.

Table 6-1: Corresponding Table Of The Decimal Number, The Binary Number And The Real Number

The decimal number	The binary number							The Corresponding real number
	Position of the bit							
	6	5	4	3	2	1	0	
0	0	0	0	0	0	0	0	0.00
1	0	0	0	0	0	0	1	0.01
2	0	0	0	0	0	1	0	0.02
3	0	0	0	0	0	1	1	0.02
4	0	0	0	0	1	0	0	0.03
5	0	0	0	0	1	0	1	0.04
6	0	0	0	0	1	1	0	0.05
7	0	0	0	0	1	1	1	0.06
8	0	0	0	1	0	0	0	0.06
…								…
126	1	1	1	1	1	1	0	0.99
127	1	1	1	1	1	1	1	1.00

1 variable is expressed by 7 bits, therefore 2 variables needs 14 bits. The gene structure is exhibited in Table 6-2.

Table 6-2: The gene structure

α_1							α_2						
Position of the bit													
13	12	11	10	9	8	7	6	5	4	3	2	1	0
0-1	0-1	0-1	0-1	0-1	0-1	0-1	0-1	0-1	0-1	0-1	0-1	0-1	0-1

C. *The flow of Algorithm*

The flow of algorithm is exhibited in Figure 6-1.

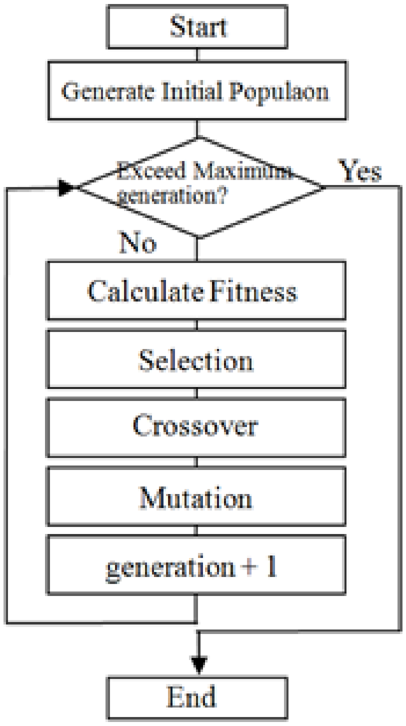


Figure 6-1: The flow of algorithm

A. *Initial Population*

Generate M initial population. Here, $M = 100$. Generate each individual so as to satisfy (22).

B. *Calculation of Fitness*

First of all, calculate forecasting value. There are 36 monthly data for each case. We use 24 data(1st to 24th) and remove trend by the method stated in section 3. Then we calculate monthly ratio by the method stated in section 4. After removing monthly trend, the method stated in section 2 is applied and Exponential Smoothing Constant with minimum variance of forecasting error is estimated.

Then 1 step forecast is executed. Thus, data is shifted to 2nd to 25th and the forecast for 26th data is executed consecutively, which finally reaches forecast of 36th data. To examine the accuracy of forecasting, variance of forecasting error is calculated for the data of 25th to 36th data. Final forecasting data is obtained by multiplying monthly ratio and trend. Variance of forecasting error is calculated by (24). Calculation of fitness is exhibited in Figure 6-2.

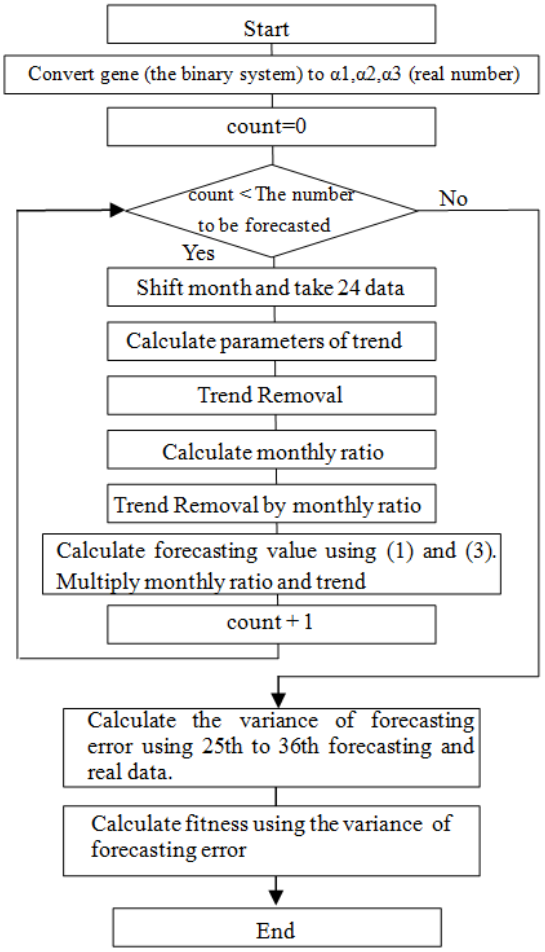

Start
Convert gene (the binary system) to α1,α2,α3 (real number)
count=0
count < The number to be forecasted
No
Yes
Shift month and take 24 data
Calculate parameters of trend
Trend Removal
Calculate monthly ratio
Trend Removal by monthly ratio
Calculate forecasting value using (1) and (3). Multiply monthly ratio and trend
count + 1
Calculate the variance of forecasting error using 25th to 36th forecasting and real data.
Calculate fitness using the variance of forecasting error
End

Figure 6-2: The flow of calculation of fitness

Scaling [22] is executed such that fitness becomes large when the variance of forecasting error becomes small. Fitness is defined as follows

$$f(\alpha_1, \alpha_2) = U - \sigma_{\varepsilon}^{2}(\alpha_1, \alpha_2) \tag{30}$$

Where U is the maximum of $\sigma_{\varepsilon}^{2}(\alpha_1, \alpha_2)$ during the past W generation. Here, W is set to be 5.

C. *Selection*

Selection is executed by the combination of the general elitist selection and the tournament selection. Elitism is executed until the number of new elites reaches the predetermined number. After that, tournament selection is executed and selected.

D. *Crossover*

Crossover is executed by the uniform crossover. Crossover rate is set as follows.

$$P_c = 0.7 \tag{31}$$

E. Mutation

Mutation rate is set as follows

$$P_m = 0.05 \tag{32}$$

Mutation is executed to each bit at the probability P_m, therefore all mutated bits in the population M becomes $P_m \times M \times 14$.

VII. Numerical Example

A. *Application to the original production data of Wheelchairs*

The data of Operating equipment and supplies for three cases (An injection device and a puncture device, A sterilized hypodermic needle and A sterilized syringe) from January 2010 to December 2012 are analyzed. These data are obtained from the Annual Report of Statistical Investigation on Statistical-Survey-on-Trends-in-Pharmaceutical-Production by Ministry of Health, Labour and Welfare in Japan. Furthermore, GA results are compared with the calculation results of all considerable cases in order to confirm the effectiveness of GA approach. First of all, graphical charts of these time series data are exhibited in Figure 7-1 - 7-3.

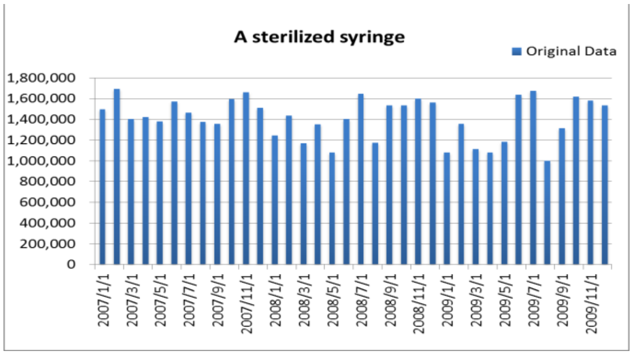


Figure 7-3: Domestic shipment data of a sterilized syringe

B. *Execution Results*

GA execution condition is exhibited in Table 7-1.

TABLE7-1: GA EXECUTION CONDITION

GA execution condition	
Population	100
Maximum Generation	50
Crossover rate	0.7
Mutation ratio	0.05
Scaling window size	5
The number of elites to retain	2
Tournament size	2

We made 10 times repetition and the maximum, average, minimum of the variance of forecasting error and the average of convergence generation are exhibited in Table 7-2 and 7-3.

The variance of forecasting error for the case monthly ratio is not used is smaller than the case monthly ratio is used in A sterilized hypodermic needle. Other cases had good results in the case monthly ratio was used.

TABLE7-2: GA EXECUTION RESULTS (MONTHLY RATIO IS NOT USED)

Food No	The variance of forecasting error			Average of convergence generation
	Maximum	Average	Minimum	
An injection device and a puncture device	551,855,685,384	504,204,854,970	497,434,740,003	15.7
A sterilized hypodermic needle	91,638,679,323	37,319,843,068	28,489,531,611	9.5
A sterilized syringe	176,511,823,650	93,652,636,003	82,555,206,063	14.1

TABLE7-3: GA EXECUTION RESULTS (MONTHLY RATIO IS USED)

Food No	The variance of forecasting error			Average of convergence generation
	Maximum	Average	Minimum	
An injection device and a puncture device	162,051,318,390	106,546,694,680	95,793,948,965	7.3
A sterilized hypodermic needle	79,422,467,024	47,074,155,352	42,493,594,397	11.1
A sterilized syringe	65,371,396,358	38,622,248,849	35,196,139,960	12.4

The minimum variance of forecasting error of GA coincides with those of the calculation of all considerable cases and it shows the theoretical solution. Although it is a rather simple problem for GA, we can confirm the effectiveness of GA approach. Further study for complex problems should be examined hereafter.

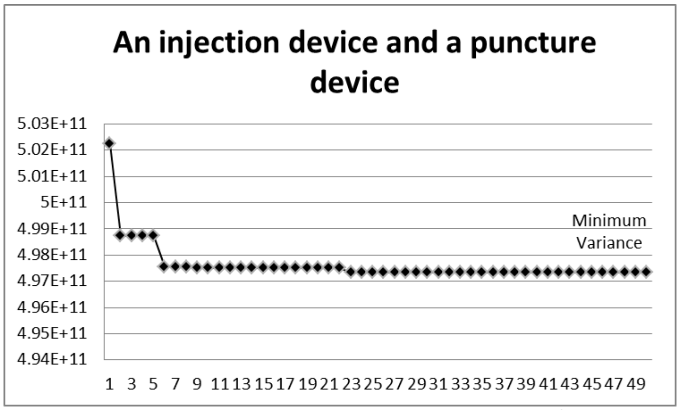


Figure7-4:Convergence Process in the case of An injection device and a puncture device (Monthly ratio is not used)

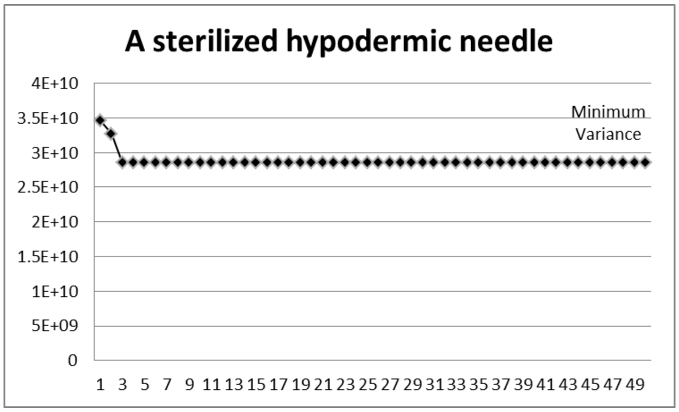


Figure7-5:Convergence Process in the case of
A sterilized hypodermic needle (Monthly ratio is used)

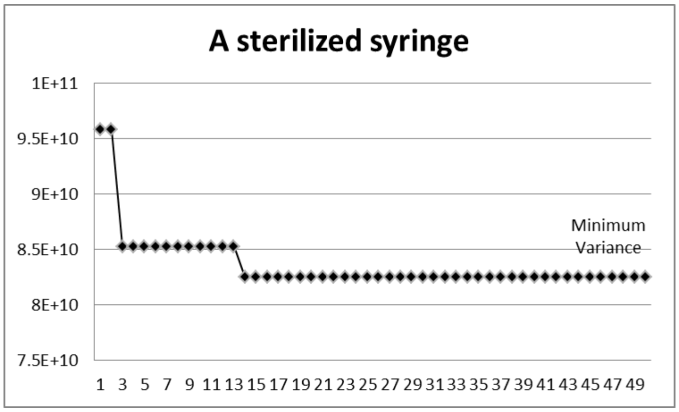


Figure7-6:Convergence Process in the case of
A sterilized syringe (Monthly ratio is not used)

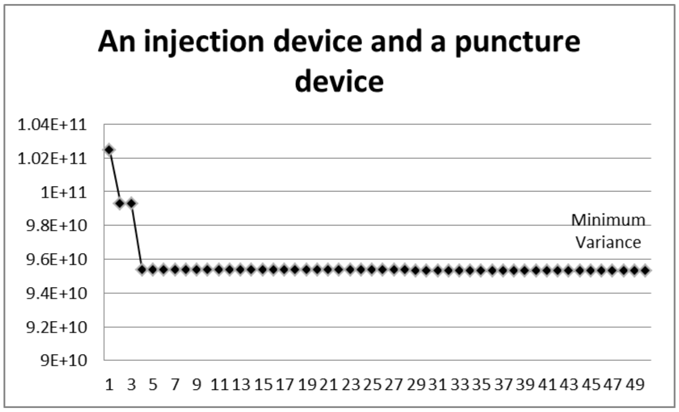


Figure7-7:Convergence Process in the case of An injection device and a puncture device
(Monthly ratio is used)

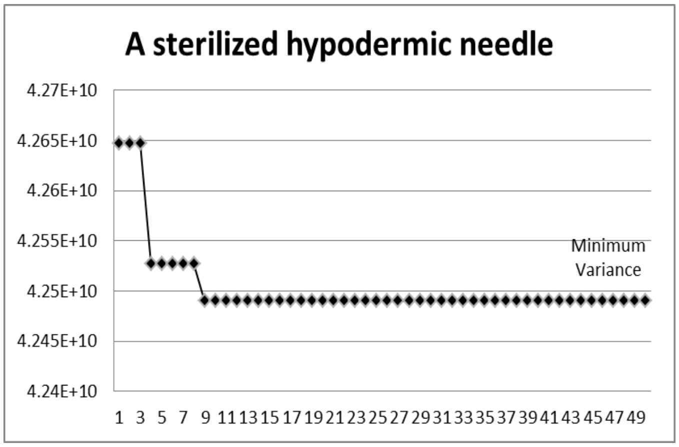


Figure7-8: Convergence Process in the case of
A sterilized hypodermic needle (Monthly ratio is not used)

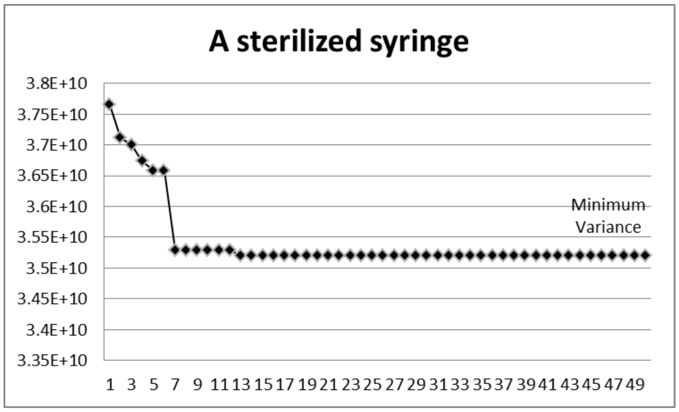


Figure7-9:Convergence Process in the case of A sterilized syringe (Monthly ratio is used)

Next, optimal weights and their genes are exhibited in Table 7-4,7-5.

TABLE7-4: OPTIMAL WEIGHTS AND THEIR GENES (MONTHLY RATIO IS NOT USED)

Data	α_1	α_2	α_3	position of the bit													
				13	12	11	10	9	8	7	6	5	4	3	2	1	0
An injection device and a puncture device	0.23	0.77	0	0	0	1	1	1	0	1	1	1	0	0	0	1	0
A sterilized hypodermic needle	0.83	0.08	0.09	1	1	0	1	0	0	1	0	0	0	1	0	1	0
A sterilized syringe	1.00	0	0	1	1	1	1	1	1	1	0	0	0	0	0	0	0

TABLE 7-5: OPTIMAL WEIGHTS AND THEIR GENES (MONTHLY RATIO IS USED)

Data	α_1	α_2	α_3	position of the bit													
				13	12	11	10	9	8	7	6	5	4	3	2	1	0
An injection device and a puncture device	0.02	0.98	0	0	0	0	0	0	1	0	1	1	1	1	1	0	1
A sterilized hypodermic needle	0	0.37	0.63	0	0	0	0	0	0	0	0	1	0	1	1	1	1
A sterilized syringe	1.00	0	0	1	1	1	1	1	1	1	0	0	0	0	0	0	0

The linear function model is best in A sterilized syringe. An injection device and a puncture device selected 1^{st}+2^{nd} order function as the best one. A sterilized hypodermic needle selected 1^{st} +2^{nd} +3^{rd} order function as the best one.

These results were same for both of “Monthly ratio is not used” and “Monthly ratio is not used”. Parameter estimation results for the trend of equation (21) using least square method are exhibited in Table 7-6 for the case of 1st to 24th data.Trend curves are exhibited in Figure 7-10 - 7-12.

TABLE 7-6: PARAMETER ESTIMATION RESULTS FOR THE TREND OF EQUATION (21)

Data	a_1	b_1	a_2	b_2	c_2	a_3	b_3	c_3	d_3
An injection device and a puncture device	8836.40	3971839.34	3616.96	-81587.61	4363676.73	466.59	-13880.33	96978.10	3954240.07
A sterilized hypodermic needle	7625.06	717466.59	369.11	-1602.74	757453.7	242.23	-8714.55	91099.12	544895.87
A sterilized syringe	-2041.09	1469255.56	1353.07	-35867.75	1615837.78	161.02	-4685.33	25756.08	1474539.34

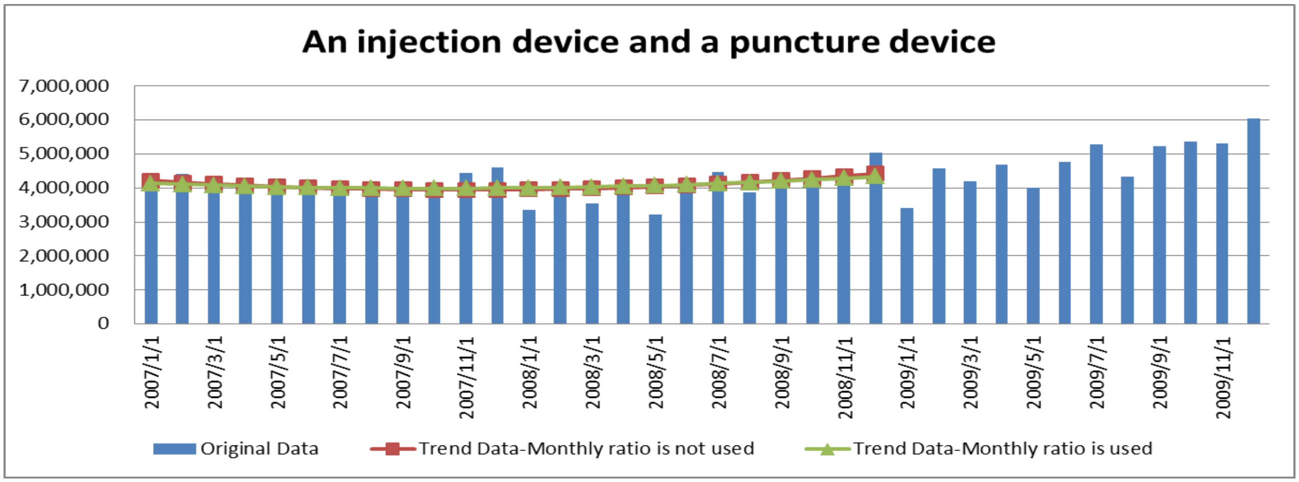


Figure7-10: Trend of An injection device and a puncture device

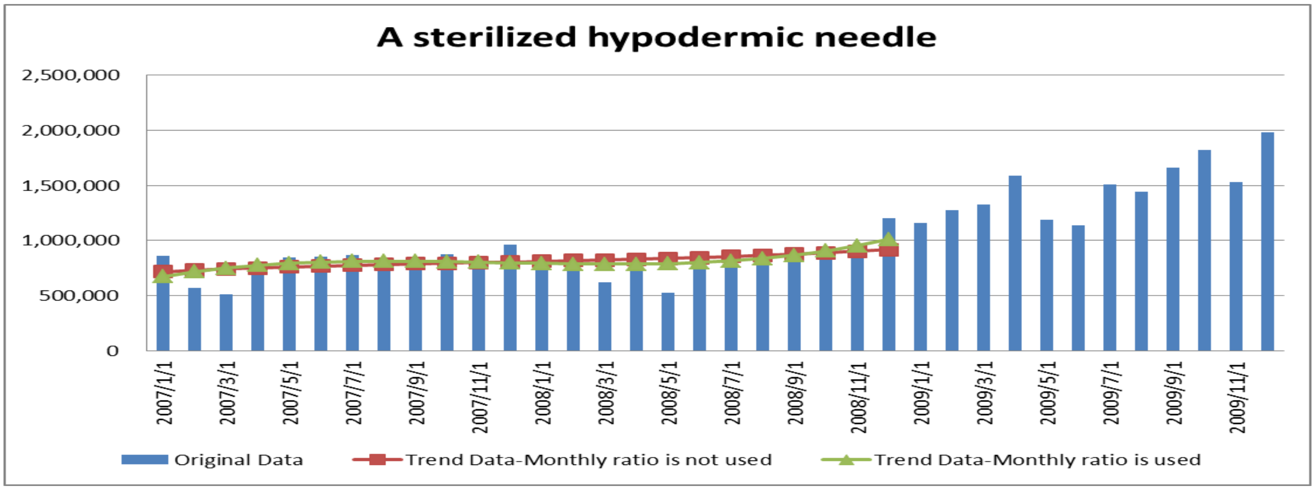


Figure7-11: Trend of A sterilized hypodermic needle

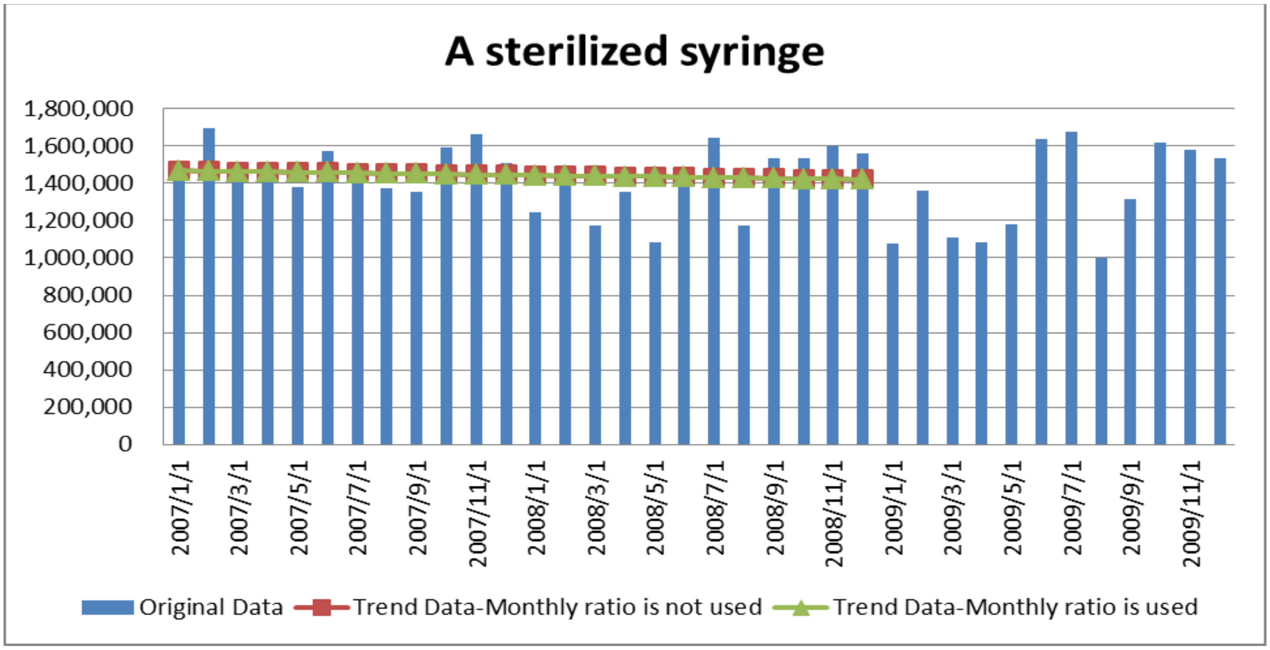


Figure7-12: Trend of A sterilized syringe

Calculation results of Monthly ratio for 1st to 24th data are exhibited in Table 7-7.

Table 7-7: PARAMETER ESTIMATION RESULT OF MONTHLY RATIO

Date.	1	2	3	4	5	6	7	8	9	10	11	12
An injection device and a puncture device	0.901	1.038	0.960	1.015	0.868	0.997	1.039	0.968	0.999	0.995	1.063	1.158
A sterilized hypodermic needle	1.146	0.900	0.735	1.072	0.866	1.033	1.094	0.984	0.993	0.993	0.986	1.198
A sterilized syringe	0.941	1.076	0.887	0.95	0.84	1.02	1.078	0.883	1.004	1.088	1.137	1.071

Estimation result of the smoothing constant of minimum variance for the 1st to 24th data are exhibited in Table 7-8, 7-9. Forecasting results are exhibited in Table 7-13 - 7-15.

Table 7-8:SMOOTHING CONSTANT OF MINIMUM VARIANCE OF EQUATION (17) (MONTHLY RATIO IS NOT USED)

Date	ρ1	α
An injection device and a puncture device	-0.147880	0.848737
A sterilized hypodermic needle	-0.342724	0.603356
A sterilized syringe	-0.499464	0.045270

Table 7-9: SMOOTHING CONSTANT OF MINIMUM VARIANCE OF EQUATION (17) (MONTHLY RATIO IS USED)

Date,	ρ1	α
An injection device and a puncture device	-0.293350	0.675822
A sterilized hypodermic needle	-0.067694	0.931993
A sterilized syringe	-0.171119	0.823554

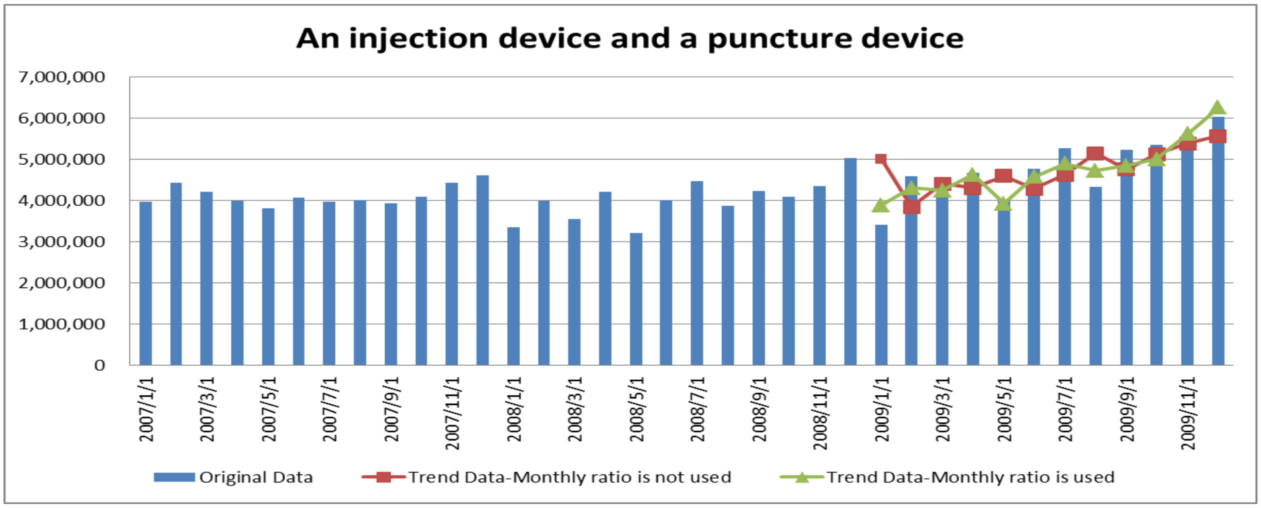


Figure 7-13: Forecasting Result of An injection device and a puncture device

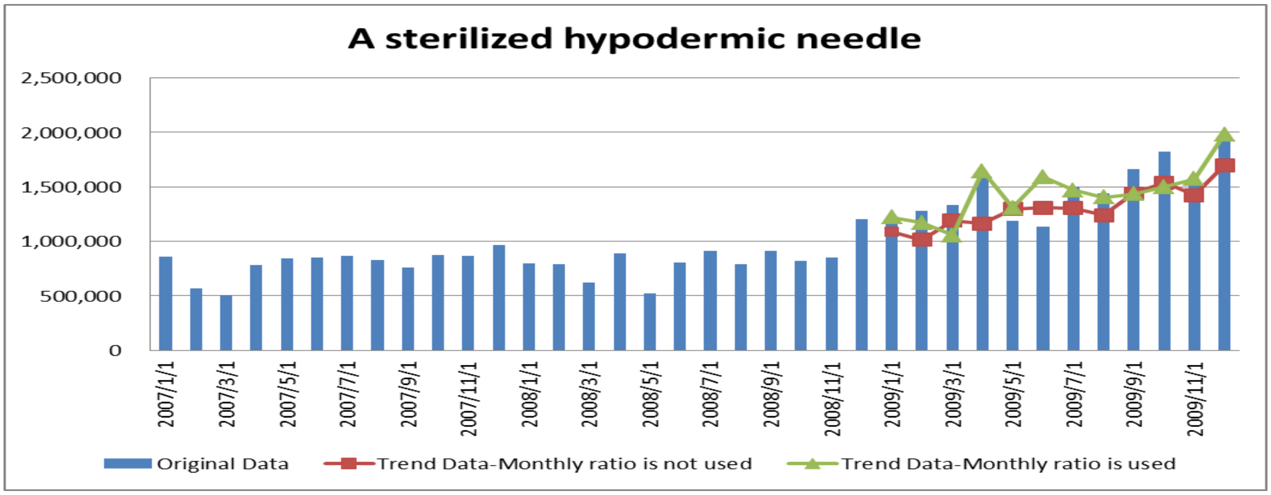


Figure 7-14: Forecasting Result of A sterilized hypodermic needle

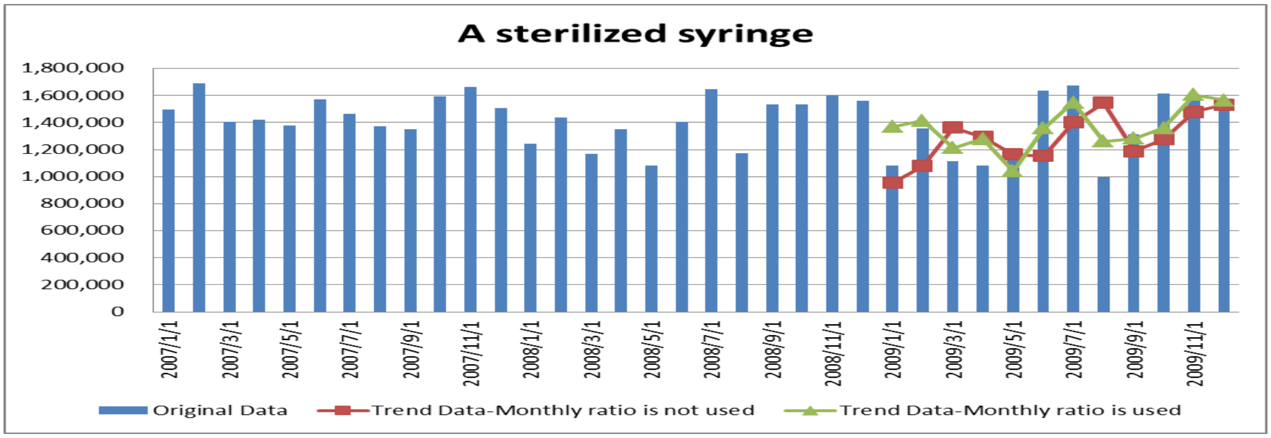


Figure 7-15: Forecasting Result of A sterilized syringe

C. Remarks

The linear function model in the case Monthly ratio was used was best for A sterilized syringe case. $1^{st}+2^{nd}$ function model in the case Monthly ratio was used was best for An injection device and a puncture device case. $1^{st}+2^{nd}$ $+3^{rd}$ function model in the case Monthly ratio was not used was best for A sterilized hypodermic needle case.

The minimum variance of forecasting error of GA coincides with those of the calculation of all considerable cases and it shows the theoretical solution. Although it is a rather simple problem for GA, we can confirm the effectiveness of GA approach. Further study for complex problems should be examined hereafter.

VIII. Conclusion

Focusing on the idea that the equation of exponential smoothing method(ESM) was equivalent to (1,1) order ARMA model equation, a new method of estimation of smoothing constant in exponential smoothing method was proposed before by us which satisfied minimum variance of forecasting error. Generally, smoothing constant was selected arbitrarily. But in this paper, we utilized above stated theoretical solution. Firstly, we made estimation of ARMA model parameter and then estimated smoothing constants. Thus theoretical solution was derived in a simple way and it might be utilized in various fields.

Furthermore, combining the trend removal method with this method, we aimed to improve forecasting accuracy. An approach to this method was executed in the following method. Trend removal by a linear function was applied to the data of Operating equipment and supplies for three cases (An injection device and a puncture device, A sterilized hypodermic needle and A sterilized syringe). The combination of linear and non-linear function was also introduced in trend removal. Genetic Algorithm was utilized to search the optimal weight for the weighting parameters of linear and non-linear function. For the comparison, monthly trend was removed after that. Theoretical solution of smoothing constant of ESM was calculated for both of the monthly trend removing data and the non monthly trend removing data. Then forecasting was executed on these data. The new method shows that it is useful for the time series that has various trend characteristics. The effectiveness of this method should be examined in various cases.

References

[1] Box Jenkins. (1994) Time Series Analysis Third Edition, Prentice Hall.

[2] R.G. Brown. (1963) Smoothing, Forecasting and Prediction of Discrete –Time Series, Prentice Hall.

[3] Hidekatsu Tokumaru et al. (1982) Analysis and Measurement –Theory and Application of Random data Handling, Baifukan Publishing.

[4] Kengo Kobayashi. (1992) Sales Forecasting for Budgeting, Chuokeizai-Sha Publishing.

[5] Peter R.Winters. (1984) Forecasting Sales by Exponentially Weighted Moving Averages, Management Science, Vol6, No.3, pp. 324-343.

[6] Katsuro Maeda. (1984) Smoothing Constant of Exponential Smoothing Method, Seikei University Report Faculty of Engineering, No.38, pp. 2477-2484.

[7] M.West and P.J.Harrison. (1989) Baysian Forecasting and Dynamic Models, Springer-Verlag, New York.

[8] Steinar Ekern. (1982) Adaptive Exponential Smoothing Revisited, Journal of the Operational Research Society, Vol. 32 pp. 775-782.

[9] F.R.Johnston. (1993) Exponentially Weighted Moving Average (EWMA) with Irregular Updating Periods, Journal of the Operational Research Society, Vol.44, No.7 pp. 711-716.

[10] Spyros Makridakis and Robeat L.Winkler. (1983) Averages of Forecasts ; Some Empirical Results, Management Science, Vol.29, No.9, pp. 987-996.

[11] Naohiro Ishii et al. (1991) Bilateral Exponential Smoothing of Time Series, Int.J.System Sci., Vol.12, No.8, pp. 997-988.

[12] Kazuhiro Takeyasu and Keiko Nagata.(2010) Estimation of Smoothing Constant of Minimum Variance with Optimal Parameters of Weight, International Journal of Computational Science Vol.4,No.5, pp. 411-425.

[13] Kazuhiro Takeyasu, Keiko Nagata, Yuki Higuchi. (2009) Estimation of Smoothing Constant of Minimum Variance And Its Application to Shipping Data With Trend Removal Method, Industrial Engineering & Management Systems (IEMS),Vol.8,No.4, pp.257-263,

[14] Kazuhiro Takeyasu, Keiko Nagata, Yui Nishisako. (2010) A Hybrid Method to Improve Forecasting Accuracy Utilizing Genetic Algorithm And Its Application to Industrial Data, NCSP'10, Honolulu,Hawaii,USA

[15] Kazuhiro Takeyasu, Keiko Nagata, Kana Takagi. (2010) Estimation of Smoothing Constant of Minimum Variance with Optimal Parameters of Weight, NCSP'10, Honolulu,Hawaii,USA

[16] Kazuhiro Takeyasu, Keiko Nagata, Tomoka Kuwahara. (2010) Estimation of Smoothing Constant of Minimum Variance Searching Optimal Parameters of Weight, NCSP'10, Honolulu,Hawaii,USA

[17] Kazuhiro Takeyasu, Keiko Nagata, Mai Ito, Yuki Higuchi (2010). A Hybrid Method to Improve Forecasting Accuracy Utilizing Genetic Algorithm, The 11th APEIMS, Melaka, Malaysia

[18] Kazuhiro Takeyasu, Keiko Nagata, Kaori Matsumura. (2011) Estimation of Smoothing Constant of Minimum Variance and Its Application to Sales Data, JAIMS, Honolulu, Hawaii, USA

[19] Hiromasa Takeyasu, Yuki Higuchi, Kazuhiro Takeyasu. (2012) A Hybrid Method to Improve Forecasting Accuracy in the Case of Bread, International Journal of Information and Communication Technology Research, Vol.2 , No.11, pp.804～812.

[20] Kazuhiro Takeyasu and Kazuko Nagao.(2008) Estimation of Smoothing Constant of Minimum Variance and its Application to Industrial Data, Industrial Engineering and Management Systems, vol.7, no. 1, pp. 44-50.

[21] Masatosi Sakawa. Masahiro Tanaka. (1995) Genetic Algorithm、Asakura Pulishing Co., Ltd.

[22] Hitoshi Iba.（2002）Genetic Algorithm、Igaku Publishing.

User-Based Interaction for Content-Based Image Retrieval by Mining User Navigation Patterns.

A.Srinagesh[1]
CSE Department, RVR & JC College of Engineering
Guntur-522019. India

Lavanya Thota[1]
CSE Department, RVR & JC College of Engineering
Guntur-522019. India.

G.P.Saradhi Varma[2]
IT Department, SRKR College of Engineering
Bhimavaram-534204, India

A.Govardhan[3]
CSE Department, JNTUH
Hyderabad-500085. India

***Abstract*—In Internet, Multimedia and Image Databases image searching is a necessity. Content-Based Image Retrieval (CBIR) is an approach for image retrieval. With User interaction included in CBIR with Relevance Feedback (RF) techniques, the results are obtained by giving more number of iterative feedbacks for large databases is not an efficient method for real-time applications. So, we propose a new approach which converges rapidly and can aptly be called as Navigation Pattern-Based Relevance Feedback *(NPRF)* with User-based interaction mode. We combined NPRF with RF techniques with three concepts viz., query Re-weighting (QR), Query Expansion (QEX) and Query Point Movement (QPM). By using, these three techniques efficient results are obtained by giving a small number of feedbacks. The efficiency of the proposed method with results is proved by calculating Precision, Recall and Evaluation measures.**

Keywords*—Image Retrieval; CBIR; Relevance Feedback; Navigation Patterns; Query Expansion; Query Reweighting; Query Point Movement.

I. INTRODUCTION

The popularity of an image retrieval system plays an important role in it's usage and application which, in turn is dependent on it's implementation. Image retrieval systems such as CBIR take a great challenge of retrieving images from a large database. Everywhere, we see the usage of images and image retrieval technique plays a vital role in different application areas like for ex: Medical Diagnosis, Military, Retail catalogs etc. There are many traditional approaches for information retrieval, but they can't satisfy the user's need to retrieve images upto a satisfactory level. CBIR techniques were firstly introduced by Rui, Hunag, and Chang, in late 1990s. There are some CBIR techniques which are search or retrieval type with browsing as a major mode of querying. CBIR is based on navigating an image collection of size 35,000 along conceptual dimensions that describes images in the collection is a very much useful method. It can also be used for intelligent image retrieval and browsing using semantic web-based techniques. All systems introduced for automatically classifying images gathered on the Web are based on the CBIR system. Another example system is art image retrieval based on user profiles is developed and it uses probabilistic support vector machines (SVM) to model user profiles. The same method is presented for automatic image annotation using cross media relevance models. Current interfaces of CBIR system describes an alternative interface based on a study of how home users use traditional ways of storing and organizing personal photo collections, but leveraging new possibilities enabled by digital media is not attempted. Some of researchers proposed approaches related to CBIR that involves multiple sources of information like text, HTML tags which are required to search for the images.

Several scenarios exist where medical practitioners can benefit from the use of these types of relevance feedback systems. Feedback functionality is to be provided for radiologists in assessing medical images, which is used in medical diagnosis. It is also useful as a clinical tool or in an academic context where students can benefit from access to similar diagnosed data. Content-based access to medical images has strong impacts for computer-aided diagnosis, evidence-based medicine. For each application, a certain GUI is composed and connected to the IRMA core hosting the database as well as the programs for feature extraction and comparison. However, several mechanisms are of major importance in every image retrieval system.

Feature extraction [12] is one of the ways to retrieve an image. Feature extraction plays a major role in CBIR systems. Mapping the image pixels into the feature space is feature extraction. By using this extracted feature we can search, index and browse the image from the stored database. This feature can be used to measure the similarity between the stored images.

Image retrieval approaches are based on the computing the similarity between the input query image and database images via Query by Example (QBE) system [9]. The problem occurred in this is extracted visual features are too diverse to be captured with the concept of query given by user. To, solve such problem in QBE system, user need to provide feedback like pick relevant images from retrieval of images iteratively, the feedback procedure is called Relevance Feedback (RF). This feedback is given up to user satisfaction with the retrieval results.

To solve problems we propose an approach called Navigation-Pattern-Based Relevance Feedback (NPRF) which is used to get efficient image retrieval quality with user interaction. CBIR combined with RF is a widely used technique to get high quality of image retrieval. Relevance feedback is a tool where end user involvement is more to improve the performance of the system. CBIR with RF methods are combined to discover Navigation Patterns i.e. user intentions taken in to account.

II. Related Work

The term Content-Based Image Retrieval (CBIR) seems to have originated with the work of Kato for the automatic retrieval of the images from a database, based on the color and shape present. Hiremath and Pujari proposed CBIR system is used to partitioning the image into tiles by using color, texture and shape features. In earlier studies for RF make use of existing machine learning techniques to achieve semantic image retrieval which includes Statistics, EM, KNN, etc. Although these were formulating the special semantic features for image retrieval, e.g., Photobook [5], Visual SEEK [7], QBIC [1], there still have imperfect descriptions for semantic features.

Relevance feedback (RF) [2], [8], [13] is used to increase the accuracy of image search process. Relevance feedback was first proposed by Rui *et. al* as an interactive tool in content-based image retrieval. PFRL methods is one of the method for relevance feedback to retrieve images are used to compute local feature relevance. By giving input query image to the system then top N similar results are display to the end user. The user needs to give feedback like select all similar images which are relevant to the query image given by end user. Here we classify the N images into clusters one is containing similar images and other containing dissimilar images, the features of two clusters are averaged. To improve the result, retain all the relevant images are selected by the user and discard the irrelevant images. These discarded images are replaced by new images from the database by comparing the feature vector of the images with the mean of the similar image cluster. This process is continued up to user retrieves relevant images.

A. Query-Point Movement

For obtaining accuracy of image retrieval is moving the query-point towards the contour of the user's preference in feature space. At each feedback QPM regards multiple positive examples as a new query point. According to user's interest query point should be close to a convex region. The space-vector formula is proposed by Rocchio is as follows:

$$\boldsymbol{Q_i = Q_{i-1} + \alpha \sum_{j=1}^{nr} \frac{R_j}{nr} - \beta \sum_{j=1}^{nir} \frac{IR_j}{nir} (1)}$$

Where,$Q_i$is the vector of the i^th query, R_j is the vector of the j^th relevant image, IR_j is the vector of j^th irrelevant image, nr is the cardinality of relevant images, nir is the cardinality of irrelevant images. There are many approaches one of them is modified version of MARS [4]. In that weighted Euclidean distance to compute the similarity between the query image and targets. The other well known study is MindReader [3], is to generalize Euclidean distance to compute targets. The modified query point of each feedback moves toward local optimal centroid.

B. Query Re-weighting:

In query reweighting if the i^th feature f_i exists in positive examples frequently, the system assigns a high degree to f_i .QR like approaches proposed by the Rui et al. [6] which convert image feature vectors to weighted feature vectors in early version of Multimedia *Analysis and Retrieval System (MARS)*. NNEW, developed by You et al. [12], the user learns query from positive and negative examples by weighting the important features. In RF approach feature weights are updated dynamically to connect low-level visual features and high-level human concepts. The search area is continuously updated by reweighting the features some targets are lost.

C. Query Expansion:

QPM and QR cannot elevate the quality of RF and cannot completely satisfy the user's interest spreading the feature space. QEX is the technique which solves the problem and gives the high quality of image retrieval. In this method user need to submit a query which captures an initial set of results, from these set of results number images should be relevant. Wu et al. [10] proposed FALCON, is designed to handle disjunctive queries within arbitrary metric spaces. Qcluster, developed by Kim and Chung [14], intends to handle the disjunctive queries by employing adaptive classification and cluster merging methods. The system expands the query based upon the terms in the selected images. This technique is used in search engines. When user has already found a set of appropriate results, then they may not desire to expand the query more.

D. Hybrid Approach:

Hybrid is another type of RF techniques; this method is used very little. Hybridized work focuses on the long-term usage log coming from various users. The greater effectiveness of the multisystem requires high computation cost due to multiple processing's. One of the hybrid RF techniques is IRRL proposed by Yin et al. [11]. The problem occurring in hybrid RF is that one cannot avoid the overhead of long iterations of feedback. Visual diversity existing in global feature space cannot be resolved with this technique. So, we are not using it in our proposed approach.

III. Our Methodology

Our proposed approach is NPRF, which integrates the discovered patterns and RF techniques to achieve effective results.

A. Outline of Navigation-Pattern Based Relevance Feedback

To solve the problems occurred in existing approaches NPRF approach is introduced. It is solution for getting high quality of image retrieval. NPRF approach is divided into two major phases those are *online image retrieval* and *offline knowledge discovery*. Each phase contains sub phases, query image is given to the system and it finds the most similar images without considering any feature vectors then it returns a set of most relevant images. The first query process is called *initial feedback.* If initial feedback is satisfy by the user then system is terminated. Then positive examples are picked up by the end user and send feedback to the image search phase by including new feature weights, new query points and user's feedback. Then by using the navigation patterns with navigation pattern-based relevance feedback search (NPRF search) is used to find the similar images. At each feedback the results are given to the end user and related browsing information is stored in the log database.

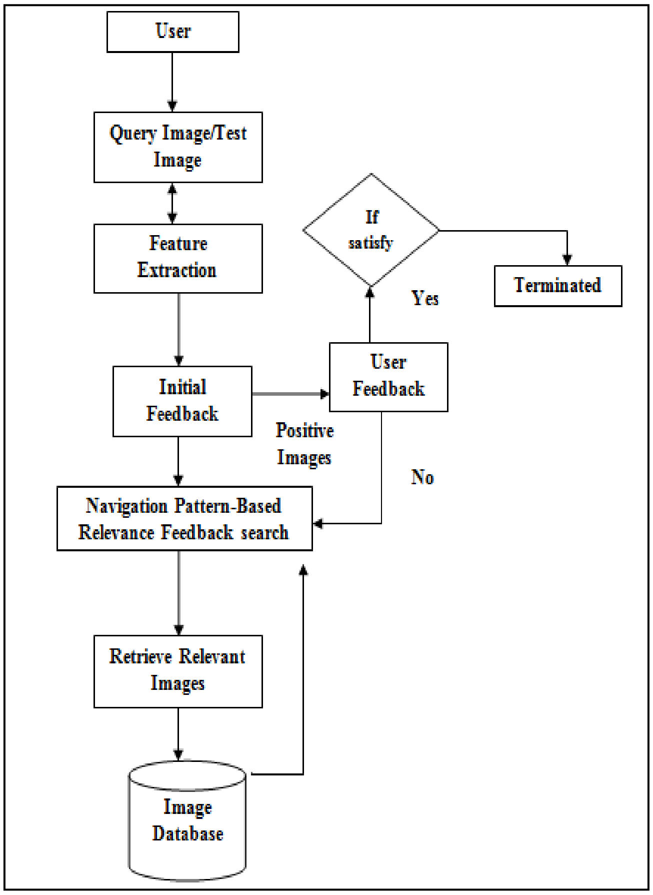


Fig. 1. Architecture of NPRF

In the above Figure, the architecture is divided in to two operations. The two steps those are *Initial query processing phase* and *image search phase*. After initial query processing phase is completed initial feedback (iteration 0) is given by this phase.

In our approach user feedback is taken into account. so, user need to give feedback as one picks positive examples, if user satisfied then system is terminated otherwise go to our proposed search NPRF search then by discovering navigation patterns relevant images are obtained.

1) *Image Retrieval*

In this phase we have sub phases those are Initial Query Processing Phase and Image search Phase.

a) Initial Query Processing Phase:

In this phase without considering any feature weights system extracts the visual features from original query image and similar images. The positive examples or good examples are picked up by the end user is called initial feedback or iteration0.

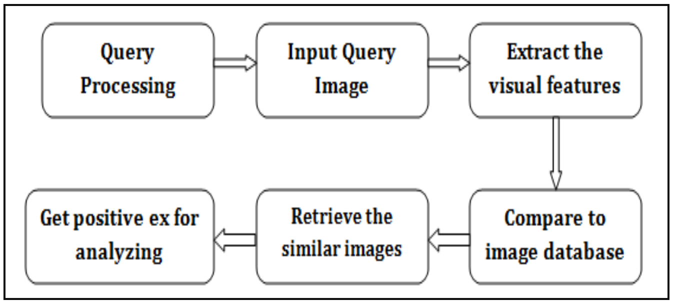

Query Processing
Input Query Image
Extract the visual features
Compare to image database
Retrieve the similar images
Get positive ex for analyzing

Fig. 2. Initial Query Processing Phase

b) Image Search Phase:

In this phase the intent is to extend one's search point to multiple search points by integrating the navigation patterns and the proposed search algorithm NPRF search. In this phase user intension is successfully implied. A new query point is generated at each feedback by using preceding positive examples. The search procedure is continuing up to user is satisfied.

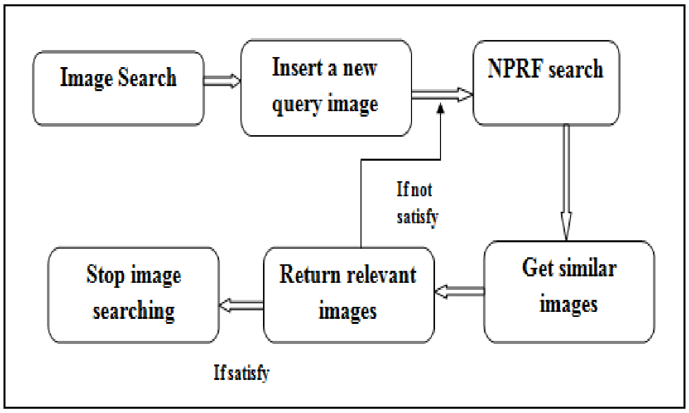


Fig. 3. Image Search Phase

B. Image Search Phase

The aim of the search strategy is to solve the problems in existing approaches; these problems result in large limitation in RF. By using RF query refinement strategies the results generated by multiple query refinement systems produce better results than individual systems. Our proposed approach NPRF Search resolves problems by using the generated navigation patterns. For the problem of existing problems like exploration convergence and redundant browsing, our proposed approach extends the search range from a query point to a number of relevant navigation paths; as a result user's interest is satisfied. The discovered navigation patterns are taken as the shortest paths to derive the efficient results in a few feedback processes. Because of high cost of navigation process for the massive image databases iterative search can be a solution. The NPRF Search algorithm can be divided as an important part of our proposed iterative solution to RF, which is combination of QEX, QP, and QR strategies.

a) Flow Chart For NPRF Approach

Flow chart of the proposed approach is shown below in fig 6. It represents a step by step procedure. Firstly a set of images are taken for which old query points are found and then user need to pick the positive examples to generate the new query points and these points are stored in a database, negative examples are appended to negative image set N[i]. At each feedback negative images are eliminated and no. of iterations are computed as we return and again go to step 1(feedback procedure) and system goes to step 2(Iteration process) where we get relevant or positive images without exceeding threshold *thrd.* Top s visual query points are generated and negative images are neglected, finally we get top k relevant examples.

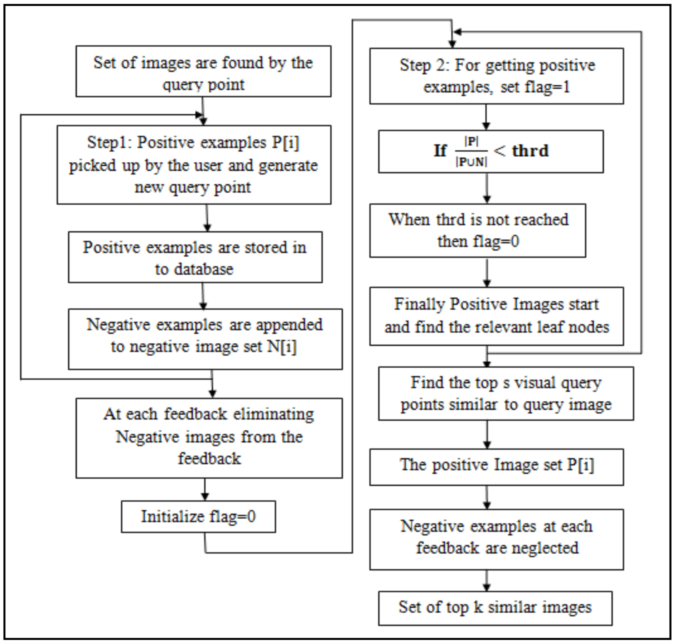

Set of images are found by the query point
Step1: Positive examples P[i] picked up by the user and generate new query point
Positive examples are stored in to database
Negative examples are appended to negative image set N[i]
At each feedback eliminating Negative images from the feedback
Initialize flag=0
Step 2: For getting positive examples, set flag=1
If $\frac{|P|}{|P \cup N|} <$ thrd
When thrd is not reached then flag=0
Finally Positive Images start and find the relevant leaf nodes
Find the top s visual query points similar to query image
The positive Image set P[i]
Negative examples at each feedback are neglected
Set of top k similar images

Fig. 4. Flow Chart of NPRF approach

A. *Algorithm NPRF Search*

NPRF Search is proposed to reach the high precision of image retrieval in a shorter query process by using the valuable navigation patterns. We explain the details of NPRF Search given below

As illustrated in NPRF Search algorithm is triggered by receiving:

a) A set of positive examples P[i] and negative examples N[i] determined by the user at the preceding feedback.

b) A set of navigation patterns.

In brief, the iterative search procedure can be decomposed into several steps as follows:

Figure a. A new query point is generated by averaging the visual features of positive examples.
Figure b. Find the similar images by determining the nearest to query image.
Figure c. Find the nearest images from the similar navigation patterns.
Figure d. Find the top s relevant visual query points from the set of the nearest images.
Figure e. Finally, the top k relevant images are returned to the user.

From the aspect of NPRF Search, step 1 can be done by QPM and steps 2-5 can be done by QEX. For QR, the feature weights are updated iteratively based on the positive examples at each feedback. The proposed NPRF Search takes advantage of Query refinement strategies and navigation patterns are used to make RF more efficiently and effectively. Without using navigation patterns, proposed search cannot reach the high quality of RF.

The goal of our approach is to satisfy each query efficiently instead of providing personalized functions for each user. By collecting a number of query transactions, most queries can be satisfy user's interests by NPRF Search. The details of the NPRF Search algorithm are described as follows:

```
Algorithm of NPRF Search:
Input: A set of positive images P[i] picked up by user, a set of
        negative set N[i];
Output:  A set of relevant images R[i];
Step 1: Generate a new query points and features $\sum_{i=1}^{n} F_i$ of
        positive images.
Step 2:  let Negative images are stored in the Negative image
        set N[i].
Step 3: Initialize flag=0;
          for each query image belongs to P[i] do
              Determine the images with the shortest distance
            to query image   $D_i = \min \sum_{i=1}^{n} (T_i - Q_i)$;
          end for
          if threshold exceeds then
              for each negative image belongs to N[i] do
                  determine the images with the shortest
               distance to negative images.
              end for
          end if

        if flag=1 then
              Find the set of visual query points with in the
            retrieved images.
        end if
        for i=1 to k do
            find the relevant image set R[i] from the database
        end for
      eliminate the negative images from retrieved images.
       return the retrieved image set R[i] of top k similar
      images.
```

Fig. 5. Algorithm Of NPRF Search

Query point generation: This operation is used to find the images may or may not use the similarity function. Modification of the query point moves query point towards the targets in search process. Assume that a set of images is found by the query points at the previous feedback. Then visual features of the positive examples P[i] picked up by the user are first averaged into a new query point. From above Fig:7, consider a set of positive examples P[i] and *d* dimensions of the ith feature$F_i = \{f_1^x, f_2^x, \dots \dots, f_d^x\}$extracted from the xth positive exampleand the positive examples are stored into the database to enhance the knowledge database. The negative examples are appended to the accumulated negative set N[i]. At each feedback, eliminating negative images from the targets can increase the precision of image retrieval significantly. In addition to generating new query points and negative images, the vectors of each feature has to be calculated to keep searching the images similar to qpnew. The feature vector for similarity computation is normalized as follows:

$$\sqrt{(X_1 - X_2)^2 + (Y_1 - Y_2)^2 + (Z_1 - Z_2)^2} \qquad \textbf{(2)}$$

Query expansion: To solve the problem of exploration convergence, this operation is used to cover all possible results by the similar patterns discovered. QEX operation first determines query seed which is nearest to each of P[i], called positive query seed, and query seed which is nearest to each of N[i], called negative query seed. From above Fig: 7 say how to find the positive and negative query seed sets.

As a result, a set of positive query seeds is selected to start the potential search paths. Because of slight loss of the information in the negative examples is also deliberated. The desired results are more precisely by discarding negative query seeds at each feedback. At each feedback there exist some query seeds which are belonging to both positive query seed set and the negative query seed set. By dropping the negative query seeds would lead to the loss of positive query seeds i.e. these dropped negative seeds may be the start of good search paths or taken for next iteration. Both positive and negative information simultaneously taken into account. If the seed gets the maximum number of negative examples or no positive example is said as bad manner, i.e., flag=0, as shown in Fig: 7.

Otherwise, flag =1 for any good manner i.e. all positive examples. By considering both positive and negative information the computation cost is more. To reduce the cost of computation, at each feedback proposed algorithm assigns a bad manner to the seed that had maximum number of negative examples, if and only if the satisfaction rate $(|P|/|P \cup N|)$ cannot reach the presetting threshold *thrd*. Finally, the good manners are the starts of the referred navigation paths to find the relevant leaf nodes.

IV. Datasets, Experimental Results

In corel database consists of seven data sets are composed of different kinds of categories. In each category contains 200 images. In our project we take a dataset consists of categories are Flowers, Animals, Mountains and Vehicles. This project is work with different databases like medical database, Wang database.

A. Datasets

For Experimentation, corel image database and web images are used in the approach. Image database which consists of corel database images, we take some of the 6 different categories from corel database those are shown in table: 1.This is also work with Wang database.

TABLE I. DATASETS

Data Set	No. of images	Category set
1	Image database (300 images)	Buses, Flowers, Mountains, Horses, Elephants and Dinosaurs
2	Wang database	Buses, Flowers, Mountains, Horses, Elephants and Dinosaurs

Experiments Datasets: By our proposed approach we have done calculations for precision, recall and evaluation measures for data set 1.

Data Set 1: In below table: 2 we have different categories those are buses, flowers, mountains, horses, Elephants and dinosaurs which are in corel database category.

TABLE II. CALCULATIONS FOR COREL DATABASE

Category	Precision	Recall	Evaluation Measure	
			b=0.5	b=2
Buses	0.416	0.086	0.8669	0.9562
Dinosaurs	1	0	1	1
Elephants	0.76	0.033	0.8892	0.9689
Flowers	1	0	1	1
Horse	0.833	0.083	0.7397	0.9150
Mountains	0.25	0.25	0.8437	0.9264

Average Precision Value is 0.709833 and Average Recall Value is0.075383.

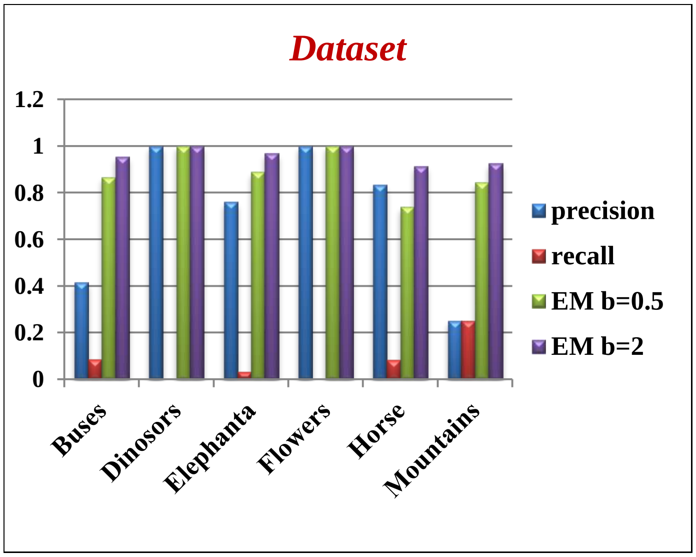


Fig. 6. Graph for Dataset Precision, recall and Evaluation measure Values

The above graph shows range of precision, recall and evaluation measure for dataset 1, where we take evaluation measure for precision values i.e. b=0.5 and b=2.

B. *Experimental Results*

Home Page:

We need to give input image and another image from image database and then click retrieve image.

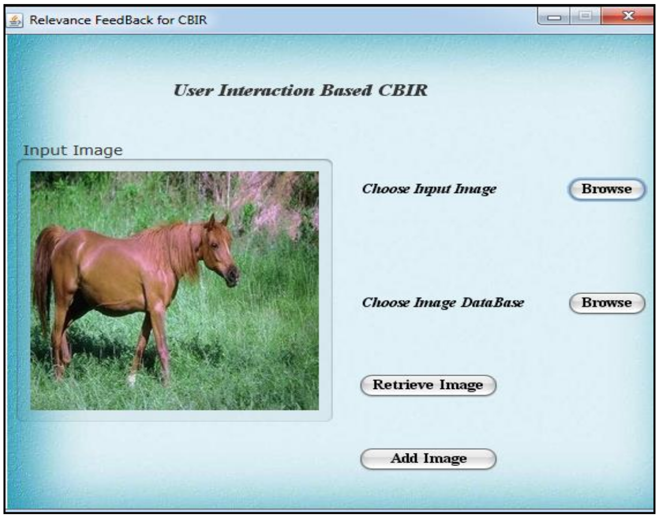
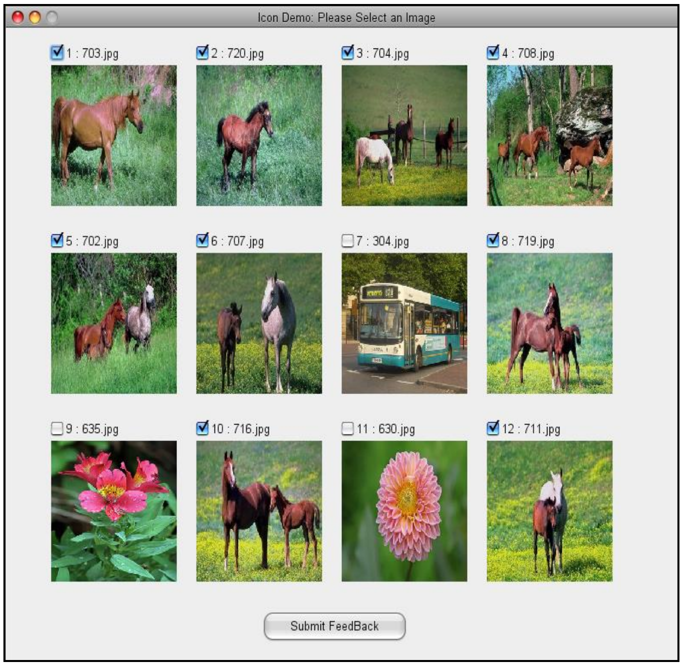


Fig. 7. Home Page

Initial Feedback:

After clicking retrieve image we get initial feedback with relevant and irrelevant images then user need to give feedback by clicking on checkbox.

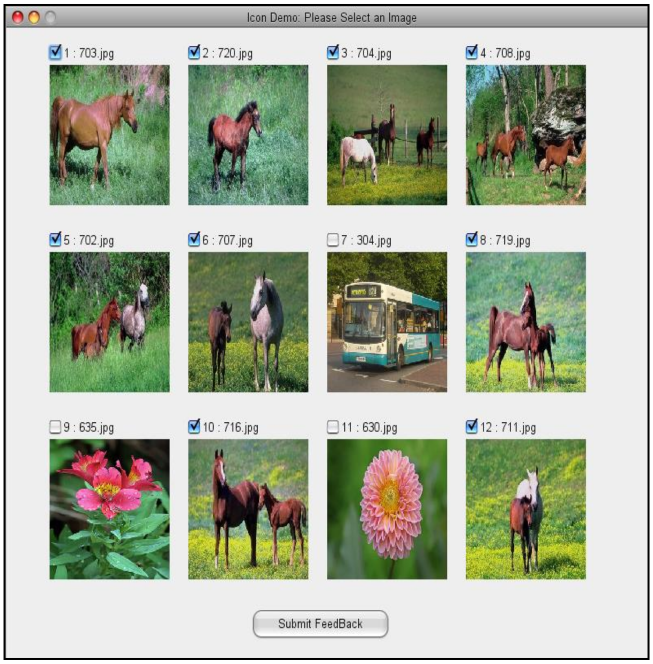


☑ 1 : 703.jpg
☑ 2 : 720.jpg
☑ 3 : 704.jpg
☑ 4 : 708.jpg
☑ 5 : 702.jpg
☑ 6 : 707.jpg
☐ 7 : 304.jpg
☑ 8 : 719.jpg
☐ 9 : 635.jpg
☑ 10 : 716.jpg
☐ 11 : 630.jpg
☑ 12 : 711.jpg

Fig. 8. Initial Feedback

User feedback:

User need to give feedback by clicking checkbox on positive image generated in initial feedback. After click submit feedback.

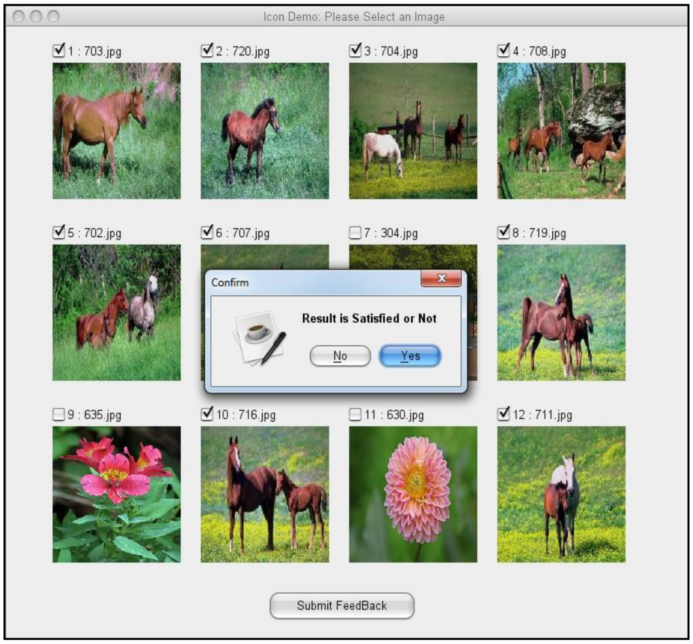


☑ 1 : 703.jpg
☑ 2 : 720.jpg
☑ 3 : 704.jpg
☑ 4 : 708.jpg
☑ 5 : 702.jpg
☑ 6 : 707.jpg
☐ 7 : 304.jpg
☑ 8 : 719.jpg
☐ 9 : 635.jpg
☑ 10 : 716.jpg
☐ 11 : 630.jpg
☑ 12 : 711.jpg

Fig. 9. User feedback

User Satisfaction:

Here user need to satisfy by result, if we click ‘yes’ then system is terminated otherwise we click ‘no’ then go to NPRF search.



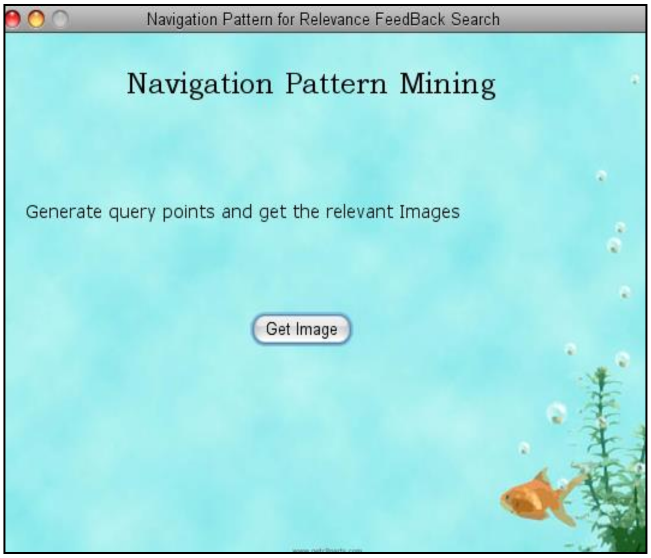
☑ 1 : 703.jpg
☑ 2 : 720.jpg
☑ 3 : 704.jpg
☑ 4 : 708.jpg
☑ 5 : 702.jpg
☑ 6 : 707.jpg
☐ 7 : 304.jpg
☑ 8 : 719.jpg
☐ 9 : 635.jpg
☑ 10 : 716.jpg
☐ 11 : 630.jpg
☑ 12 : 711.jpg

Result is Satisfied or Not

Fig. 10. User satisfaction

NPRF Search:

In this search new query points are generated and by using NPRF algorithm it produce relevant images by clicking get image button and finally we get relevant images which are relevant to input image.



Navigation Pattern for Relevance FeedBack Search

Navigation Pattern Mining

Generate query points and get the relevant Images

Fig. 11. NPRF search

Result:

Final result is produced which are relevant to query image given by the user.

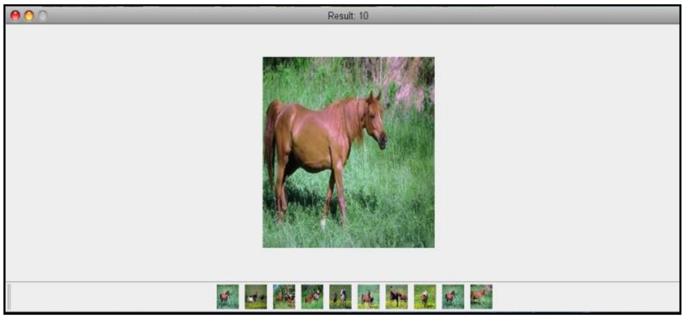
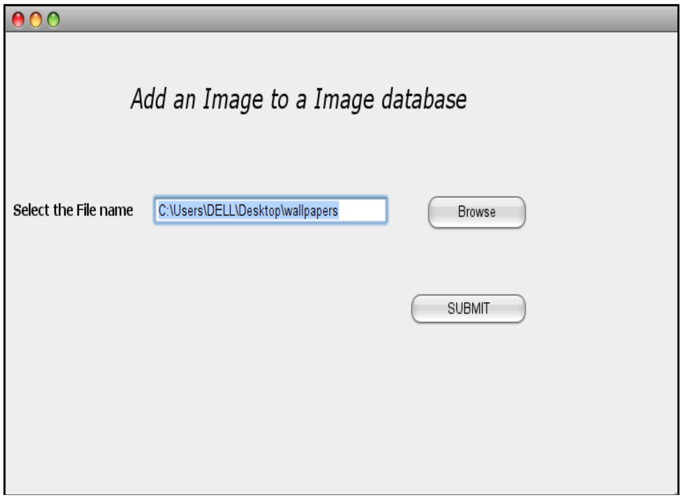


Fig. 12. Final result

Add Image:

Any image from the user's system is added to the database and one can retrieve image from that database.



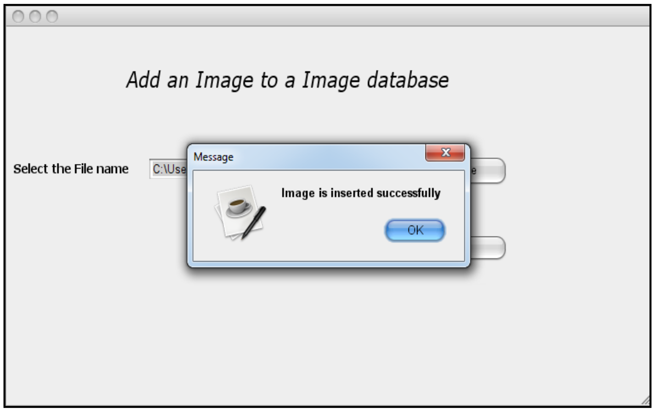
Add an Image to a Image database

Select the File name: C:\Users\DELL\Desktop\wallpapers

Fig. 13. Add Image

Insert image:

If image is inserted into database then we get message box i.e. image is inserted successfully then click "ok" button.



Add an Image to a Image database

Select the File name: C:\Use

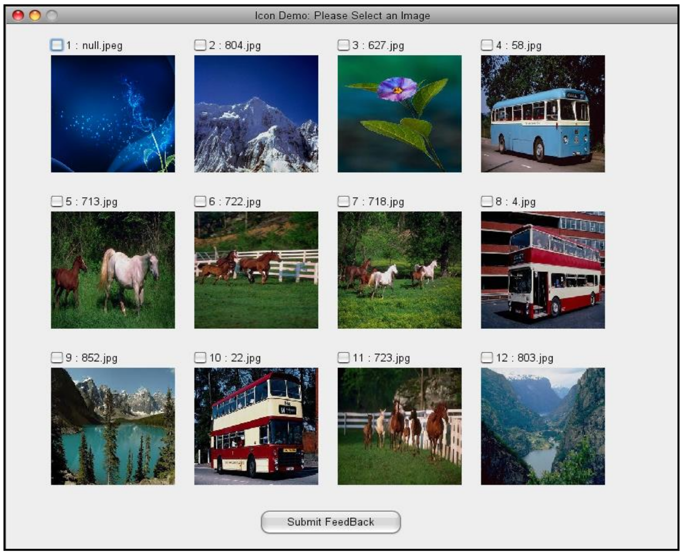
Message

Image is inserted successfully

Fig. 14. Insert Image

Retrieving Image:

Added image is to our database is "null.jpg" image which is not in our database is c:\users\DELL\Desktop\wallpapers is retrieved for next retrieval.



Icon Demo: Please Select an Image

Fig. 15. Retrieving Image

V. CONCLUSION

To solve the problem of long series of interaction with the user, the iteration technique included in CBIR combined with RF formulates into a novel method. So we have proposed a new approach called NPRF- *navigation pattern based relevance feedback*. The main aim of this approach is to get efficient results coupled with getting high quality of image retrieval with optimal or few feedbacks. Our approach will satisfy the user's intention from a long term search activity or browsing. In NPRF approach, we satisfy the user's intention by merging three query refinement strategies QR, QEX and QPM. As a result, problems occurred in existing systems like redundant browsing and visual diversity are solved. The experimental results of the proposed approach are evaluated by using precision, recall measures.

REFERENCES

[1] M.D. Flickner, H. Sawhney, W. Niblack, J. Ashley, Q. Huang, B.Dom, M. Gorkani, J. Hafner, D. Lee, D. Steele, and P. Yanker,"Query by Image and Video Content: The QBIC System," Computer, vol. 28, no. 9, pp. 23-32, Sept. 1995.

[2] D. Harman, "Relevance Feedback Revisited," Proc. 15th Ann. Int'l ACM SIGIR Conf. Research and Development in Information Retrieval, pp. 1-10, 1992.

[3] Y. Ishikawa, R. Subramanya, and C. Faloutsos, "MindReader: Querying Databases through Multiple Examples," Proc. 24th Int'l Conf. Very Large Data Bases (VLDB), pp. 218-227, 1998.

[4] K. Porkaew, K. Chakrabarti, and S. Mehrotra, "Query Refinement for Multimedia Similarity Retrieval in MARS," Proc. ACM Int'l Multimedia Conf. (ACMMM), pp. 235-238, 1999.

[5] A. Pentalnd, R.W. Picard, and S. Sclaroff, "Photobook: Content-Based Manipulation of Image Databases," Int'l J. Computer Vision (IJCV), vol. 18, no. 3, pp. 233-254, June 1996.

[6] Y. Rui, T. Huang, and S. Mehrotra, "Content-Based Image Retrieval with Relevance Feedback in MARS," Proc. IEEE Int'l Conf. Image Processing, pp. 815-818, Oct. 1997.

[7] J.R. Smith and S.F. Chang, "VisualSEEK: A Fully Automated Content-Based Image Query System," Proc. ACM Multimedia Conf., Nov. 1996.

[8] G. Salton and C. Buckley, "Improving Retrieval Performance by Relevance Feedback," J. Am. Soc. Information Science, vol. 41, no. 4, pp. 288-297, 1990.

[9] K. Vu, K.A. Hua, and N. Jiang, "Improving Image Retrieval Effectiveness in Query-by-Example Environment," Proc. 2003 ACM Symp. Applied Computing, pp. 774-781, 2003.

[10] L. Wu, C. Faloutsos, K. Sycara, and T.R. Payne, "FALCON: Feedback Adaptive Loop for Content-Based Retrieval," Proc. 26th Int'l Conf. Very Large Data Bases (VLDB), pp. 297-306, 2000.

[11] P.Y. Yin, B. Bhanu, K.C. Chang, and A. Dong, "Integrating Relevance Feedback Techniques for Image Retrieval Using Reinforcement Learning," IEEE Trans. Pattern Analysis and Machine Intelligence, vol. 27, no. 10, pp. 1536-1551, Oct. 2005.

[12] A.Sri Nagesh, Dr.G.P.S.Varma, Dr A Govardhan, "An Improved iterative Watershed and Morphological Transformation Techniques for Segmentation of Microarray Images" , IJCA Special Issue on "Computer Aided Soft Computing Techniques for Imaging and Biomedical Applications" CASCT, Volume 2, 2010.PP 77-87,August 2010.

[13] C.Gonzalez, R.E.Woods, Digital Image Processing, third ed., Prentice Hall, 2007

[14] H. You, E. Chang, and B. Li, "NNEW: Nearest Neighbor Expansion by Weighting in Image Database Retrieval," Proc. IEEE Int'l Conf. Multimedia and Expo, pp. 245-248, Aug. 2001.

[15] X.S. Zhou and T.S. Huang, "Relevance Feedback for Image Retrieval: A Comprehensive Review," Multimedia Systems, vol. 8, no. 6, pp. 536-544, Apr. 2003.

[16] D.H. Kim and C.W. Chung, "Qcluster: Relevance Feedback Using Adaptive Clustering for Content-Based Image Retrieval," Proc. ACM SIGMOD, pp. 599-610, 2003.

[17] Petrakis E. and Faloutsos A., "Similarity searching in medical image databases," Journal of IEEE Transaction on Knowledge and Data Engineering, vol. 9, pp435-447, 1997.

[18] A.Sri Nagesh, Dr.G.P.Saradhi Varma, Dr.A.Govardhan & Dr.B.Raveendra Babu, "An Analysis and Comparison of Quality Index Using Clustering Techniques for Spot Detection in Noisy Microarray Images". International Journal of Image Processing (IJIP), pp 504-511, Volume (5): Issue (4): 2011, CSC Journals, Kuala Lumpur, Malaysia, ISSN: 1985-2304.

[19] Dr. Fuhui Long, Dr. Hongjiang Zhang and Prof. David Dagan Feng, "fundamentals of content-based image retrieval.J. Weszka, C. Dyer, and A. Rosenfeld. A comparative study of texture measures for terrain classification. IEEE Transactions on Systems, Man and Cybernetics, 6(4):269– 285, 1976.

[20] Remco C. Veltkamp and Mirela Tanase, "Content Based Image Retrieval Systems: A Survey," International Journal of Engineering Science and Technology, Vol. 20, No. 5, pp. 1-62, 2002.

[21] Armitage, L and Enser, P G B (1997) "Analysis of user need in image archives." Journal of Information Science, 23(4), 287-299.

A Novel Expert System for Building House Cost Estimation: Design, Implementation, and Evaluation

Yasser A. Nada
Chairman of Department of computer science,
Faculty of Computers and Information Technology,
Taif University - K S A,

***Abstract*—This paper introduces an expert system which demonstrates a new method for accurate estimation of building house cost. This system is simple and decreases the time, the effort, and the money of its beneficiaries. In addition, design and implementation of the proposed expert system are introduced. CLIPS 6.0 and C# are used in implementation phase. Also, this expert system is programmed to be in a standalone package with a platform independency. Furthermore, the developed expert system is tested under several real cases. Finally, an initial evaluation of this expert system is carried out and a positive feedback is received from user's samples, which makes it robust and efficient.**

Keywords— Expert System; Building House; CLIPS

I. Introduction

An expert system is a computer program designed to simulate the problem-solving behavior of a human who is an expert in a narrow domain or discipline. Expert Systems (ES), also called Knowledge Based System (KBS), are computer application programs that take the knowledge of one or more human experts in a field and computerize it so that it is readily available for use. Expert System makes easier for user to identify the describe symptoms like image bases or textual bases information as it is very difficult to describe in words. It can also be integrated with textual database which can be used for explanation purposes of basic terms and operations to confirm and to reach conclusion in some situations [1].

As a branch of artificial intelligence, an expert system has been widely used. An expert system shell greatly improves the efficiency of the construction of an expert system [2]. Become a computer systems mechanism profound impact on our daily lives as we see every day research and new projects for the use of computers to make life easier and save the experience, and ease the pressure borne by the people? The paper is a mix of the latest techniques presented in this section of the computer science related systems expert systems and decision support, the paper provides scientific material distinguished and easy for the user in the field of architecture in terms of the ability to set the vision and the perception of urban are easy and available, where the idea is based on the establishment of an expert system alternative to the architect be helpful to them in calculating the cost of the construction according to data entered from (The land area, the site of the Earth And,.. etc.), which provides immediate support to customers and make decisions based on information obtained by them and contained them within the scope of existing knowledge has it .

An expert system's knowledge base is traditionally encoded as a set of domain-specific rules. These rules are generally implications of the form:

IF a1 Ù a2 Ù … Ù ak THEN ak+1 Ù ak+2 Ù … Ù an, where the ai's are logical statements that are relevant to the system's problem domain. For example, in the context of soil science, the rule:

IF a soil is sandy and the level of humus is high THEN the soil is compact

The development of expert system is implemented in CLIPS programming environment (C Language Integrated Production System) [3,4,5]. This programming tool is designed to facilitate the development of software to model human knowledge or expertise. CLIPS program is used by reason of the flexibility, the expandability and the low cost.

The outline of the paper is as follows. Section2 problem recognition. Section3 presents the basic of building a house. Section 4 knowledge representation. Section 5 tree knowledge section 6 diagnosis process finally, working example and summarizes this paper.

II. Problem Recognition

We need to build expert system to presents the design and development of an expert system for Account the Cost of Building House (ACBH). To distribute human expertise in this science.

III. The Basic of Building a House

- Choose a place of building a house
- Settlement of the land - soil quality
- Construction area
- Foundations and pillars.
- Types of foundations
- Finished Construction
- Types of buildings
- Types of fossils
- Internal planning for the home
- Determine the labor

-The numbers and types of housing required during the next twenty years in the Kingdom

IV. Knowledge Representation

The key problem is to find a KR (and a supporting reasoning system) that can make the inferences your application needs in time, that is, within the resource constraints appropriate to the problem at hand. This tension between the kinds of inferences and application "needs" and what counts as "in time" along with the cost to generate the representation itself makes knowledge representation engineering interesting.

There are representation techniques such as frames, rules, tagging, and semantic networks which have originated from theories of human information processing. Since knowledge is used to achieve intelligent behavior, the fundamental goal of knowledge representation is to represent knowledge in a manner as to facilitate inference (i.e. drawing conclusions) from knowledge [6,7].

Knowledge bases can be represented by production rules. These rules consist of a condition or premise followed by an action or conclusion (IF condition...THEN action).

For example

```
 If  Land Area 400
       and work type foundation
       and  the number of floors 1
       and the number of room 4
       and the size of water tank small
  Then cost 1,11000 SR
```

TABLE 1

Percent%	Number	Types of housing
4,23	13,812	Villa
8,04	26,267	Role at Villa
29,8	97,438	Apartment
35,77	116,841	Duplex
22,11	72,223	Mtlasq
	326,581	Total

A production rule system consists of

1- A set of rules.
2- Working memory that stores temporary data.
3- A forward or backward chaining inference engine.

Simple Examples of Represent Rules for Expert System:

FACT1	cost of the finishing normal1 equal 215,312SR
FACT2	cost of the finishing excellent equal 4,50342SR
FACT4	building foundations1 equal 252,684 SR
FACT5	building foundations2 equal 4,45469 SR
FACT 6	Land area480
FACT 7	Land area400
FACT 8	the number of floor 2
FACT 9	the number of floor 1
FACT 10	water tank small
FACT 11	water tank medium
FACT 12	driver room NO
FACT 13	driver room YES

RULE 1

```
If the land area 480
        and cost of   building foundations1
        and the cost of finishing normal1
        and the number of floor 2
 Then the overall cost744,279 SR
```

RULE 2

```
If the land area400
        and the number of floor 1
        and water tank small
Then the cost of building foundations 6,50432 SR
```

RULE 3

```
If the Finishing Normal
     and driver room NO
     and the number of floor 1
 Then cost of the finishing 4,34762 SR
```

To prove the conclusion "the overall cost744, 279 SR" inference engine must prove all condition that leading to this conclusion. Condition can be found from asking user or from another Rule because this condition is conclusion in Rule.

V. TREE KNOWLEDGE

A decision tree (or tree diagram) is a decision support tool that uses a tree-like graph or model of decisions and their possible consequences, including chance event outcomes, resource costs. Decision trees are commonly used in operations research, specifically in decision analysis, to help identify a strategy most likely to reach a goal. Another use of decision trees is as a descriptive means for calculating conditional probabilities.

We give the example for land area 480 and 400 m square.

First: land area 400 m2

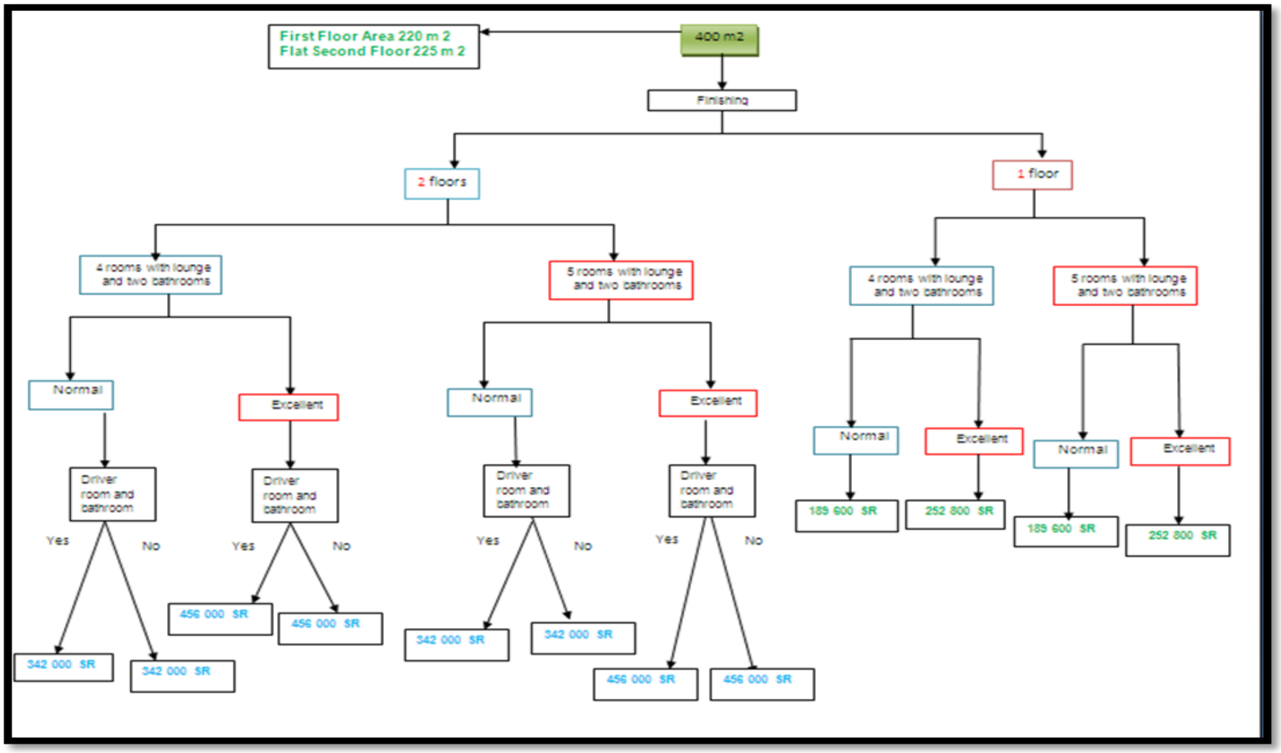


Fig. 1. Tree Knowledge - Foundation

A. Foundational

There are two options to the user in the Foundational, one or two floors.

If chosen for one floor, there are two cases before him to be the internal planning of the house 4 or 5 rooms with kitchen and two bathrooms.

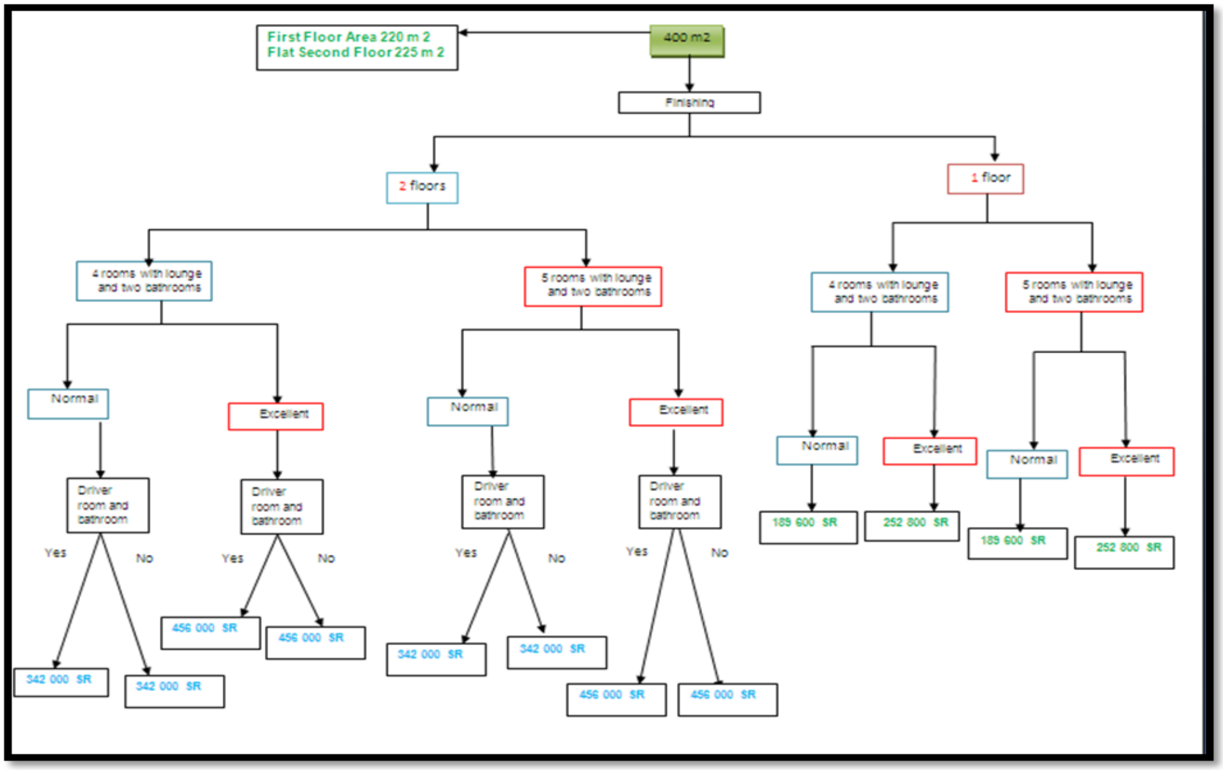


Fig. 2. Tree Knowledge – Finishing

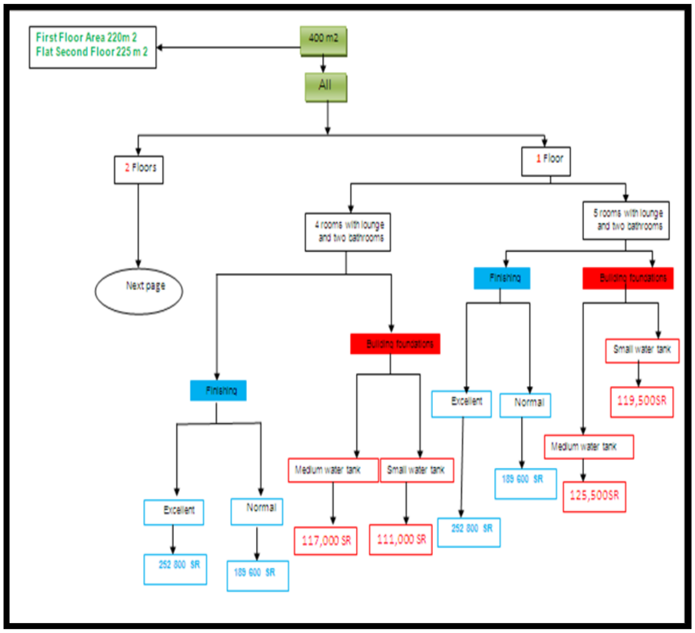


Fig. 3. Tree Knowledge – all- One

Then it will determine the size of tank water.

In the case of the user's choice 2 floors, will pass the same the previous options on one floor, but will start a driver with a bathroom extension, Show this explanation in figure (1).

B. Finishing

If the user chooses the option of calculating the cost of finishing, also given the choice between 1 or 2 floors. Then the internal planning 4 or 5 with kitchen and 2 bathrooms.

We have 2 type of finishing Normal , Excellent every Each type of them has a different cost from one another and are calculated cost them m², Show this explanation in figure (2).

C. ALL

The user select choice All this mean (Foundational and finishing together), then calculate the cost at the same way previous,

Show this explanation in the following figures (3) and (4).

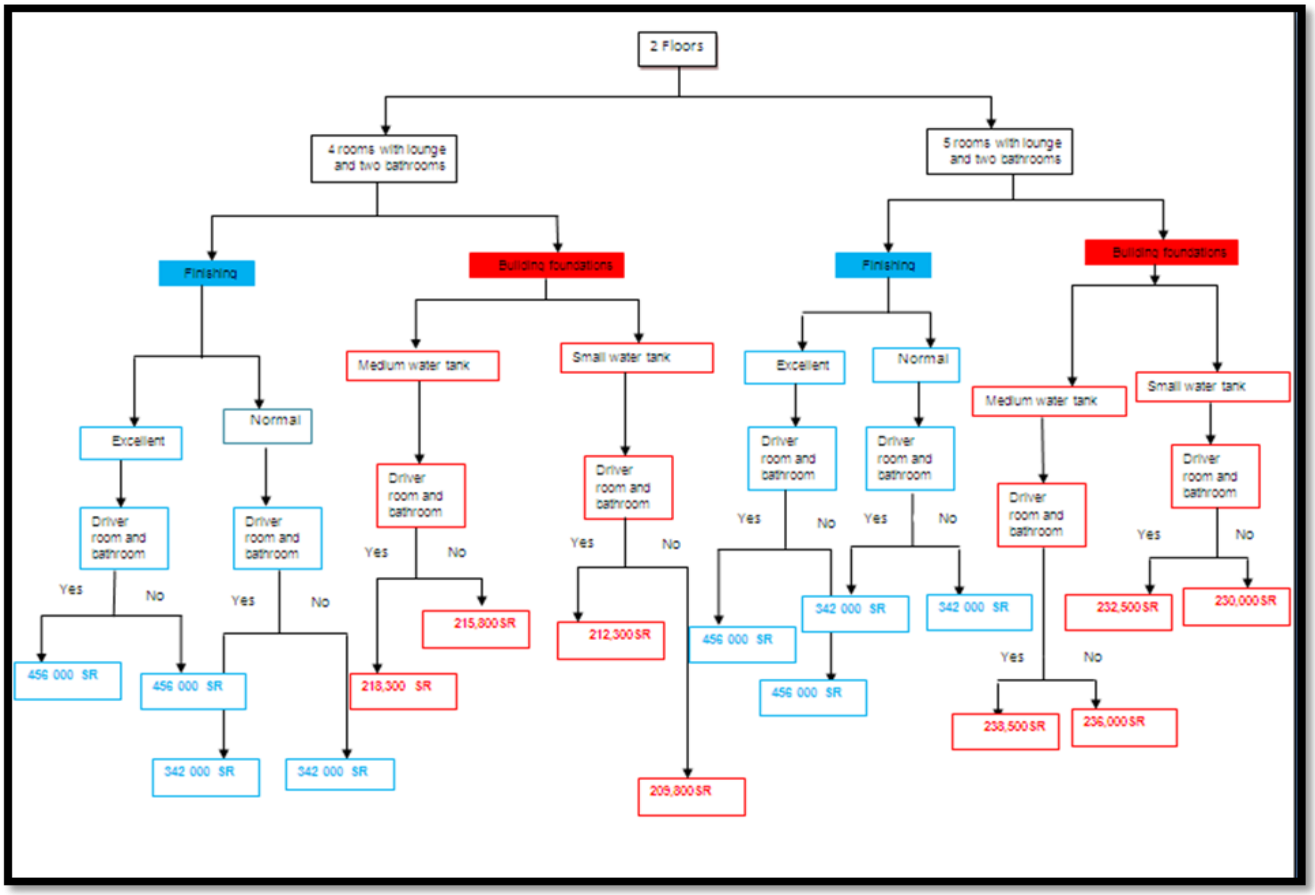


Fig. 4. Tree Knowledge – all- Two floors

VI. DIAGNOSIS PROCESS

The expert system software adopted C# to deal with the preparatory work, including the maintenance and management. All the status parameters, status values and solutions were obtained from the database access through C#, then past to CLIPS through interface functions that between C# and CLIPS, lastly diagnosed by the program that was built by CLIPS, and meanwhile, the inference information and result were past to and displayed on the interface module, which was programmed by C#.

Expert System Shell

CLIPS keep in memory a fact list, a rule list, and an agenda with activations of rules. Facts in CLIPS are simple expressions consisting of fields in parentheses. Groups of facts in CLIPS, usually follow a fact-template, so that to be easy to organize them and thus design simple rules that apply to them. Our expert system contains 100 CLIPS rules. Below, we present the rules for (ACBH).

Working Example

We can represent this rule using our representation as follow:

Some instruction in the Clips

```
       defrule work_type_process
         =>
     (printout t "foundation, finish, all" crlf)
        (bind ?answer (read idata))
        (assert(work_type ?answer))
        (printout t ?work_type)
              (defrule area_process
        =>
       (printout t "land area 400 or 480?" crlf)
       (bind ?answer (read idata))
       (assert(area ?answer)))
        (defrule f1
        (work_type foundation)(area 480)(floor one)(rooms
five )(small tank)

        =>

        (printout odata "122342" crlf))

       (defrule f2

      (work_type foundation)(area 480) (floor one)(rooms
five)(medium tank)

       => (printout odata "128342" crlf))
```

Figures 5,6,7,8,9 and 10 present some samples from the proposed expert system forms and menus.

After Execution

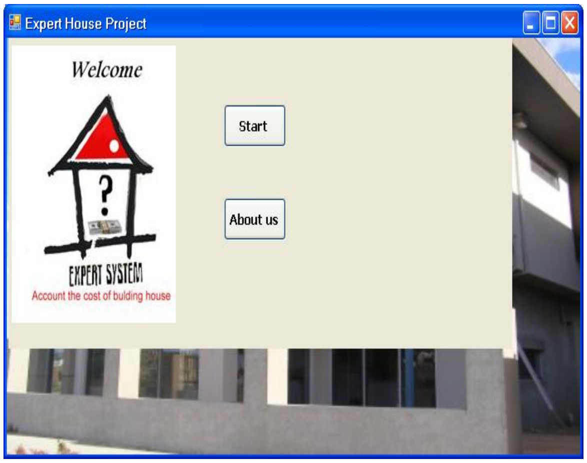
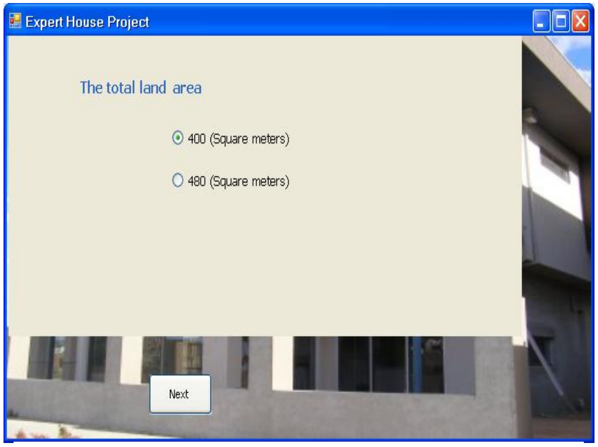


Fig.5. Start-up program

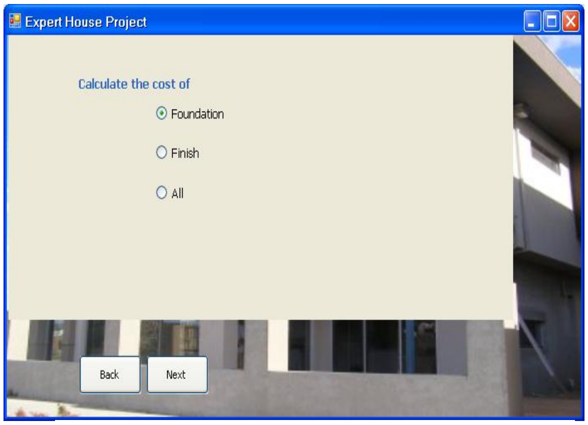


The total land area

● 400 (Square meters)

○ 480 (Square meters)

Fig.6. Second screen chose the land area



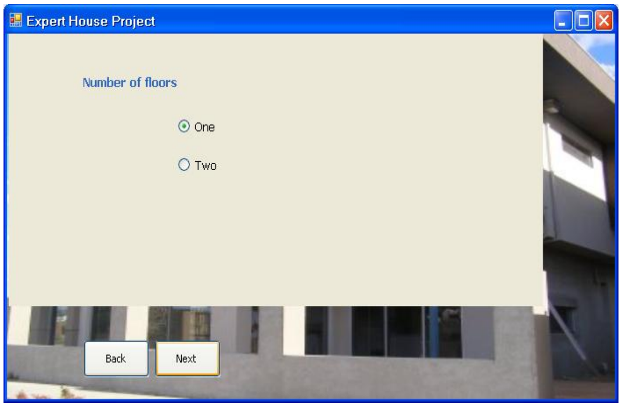
Calculate the cost of

● Foundation

○ Finish

○ All

Fig.7. choose the type of work

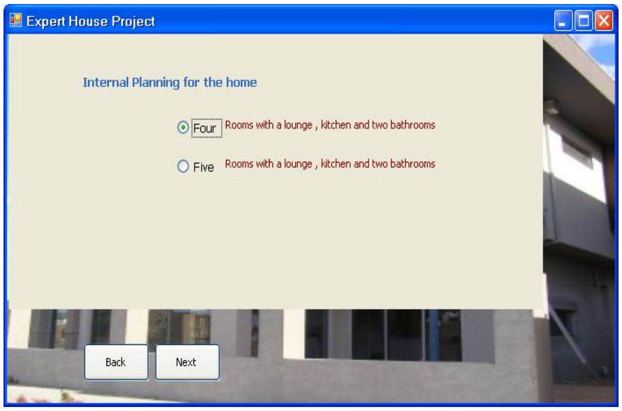


Number of floors

● One

○ Two

Fig.8. choose the number of floors

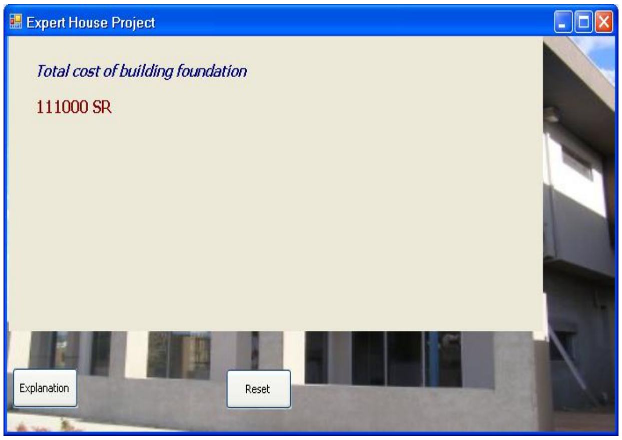


Internal Planning for the home

● Four Rooms with a lounge , kitchen and two bathrooms

○ Five Rooms with a lounge , kitchen and two bathrooms

Fig.9. choose the number of rooms



Total cost of building foundation

111000 SR

Fig.10. the result of cost SR

VII. CONCLUSIONS

In this paper, the design of an expert system for estimating the Cost of building house is introduced. The expert system is implemented using Clips to build a knowledge base and C # to design a foreground interface. The developed expert system interface is used to receive information from users and handle it under several cases. Accordingly, it returns an accurate estimation to the user. The proposed expert system is included in one executable standalone package. In addition, the proposed expert system test proves that it simple, accurate, powerful, and flexible.

REFERENCES

[1] Khumukcham R., Shikhar Kr. S, "JESS Based Expert System Architecture For Diagnosis Of Rice Plant Diseases: Design And Prototype Development", 2013 4th International Conference on Intelligent Systems, Modeling and Simulation, P 674-676, 2013.

[2] Qi Ming Cui, Sheng Wei Liu, Rong Xing Liu, and Zhi Rai Wang,"Application Research of the Production System type Expert System Shell Pro/3 under Smart Grid", IEEE PES ISGT ASIA 2012.

[3] Giarranto J. C., "CLIPS User's Guide", Version 6.22, 1998.

[4] Jackson, P., "Introduction to Expert Systems", Harlow, England: Addison Wesley Longman. Third Edition, 1999.

[5] Da-peng Tan, Shi-ming Ji Shu-ting Chen, "Continuous Casting Slag Detection Expert System Based on CLIPS", 2010 Ninth International Symposium on Distributed Computing and Applications to Business, Engineering and Science, IEEE, 2010.

[6] George F. Luger and William, A. Stubblefield; " Artificial Intelligence and the Design of Expert Systems", The Benjamin/ Cummings publishing Co., Inc., 1989.

[7] Keith D., "The Essence of Expert Systems", Prentice Hall, 2000.

AUTHORS BIOGRAPHY

DR. Yasser Ahmed Nada Was born in Ismailia, Egypt, in 1968. He received the BSc degree in pure Mathematics and Computer Sciences in 1989 and MSc degree for his work in computer science in 2003, all from the Faculty of Science, Suez Canal University, Egypt. In 2007, he received his Ph.D. in Computer Science from the Faculty of Science, Suez Canal University, Egypt. From September 2007 until now, he worked as Assistant Professor of computer science. Chair, Department of computer science, Faculty of Computers and Information Technology, Taif University, KSA. His research interests include Expert Systems, Artificial Intelligence, Object Oriented Programming, Computer Vision, and Genetic.

On Integrating Mobile Applications into the Digital Forensic Investigative Process

April Tanner, Ph.D.
Department of Computer Science
Jackson State University
Jackson, USA

Soniael Duncan
Department of Computer Science
Jackson State University
Jackson, USA

Abstract— **What if a tool existed that allowed digital forensic investigators to create their own apps that would assist them with the evidence identification and collection process at crime scenes? First responders are responsible for ensuring that digital evidence is examined in such a way that the integrity of the evidence is not jeopardized. Furthermore, they play a pivotal part in preserving evidence during the collection of evidence at the crime scene and transport to the laboratory. This paper proposes the development of a mobile application that can be developed for or created by a first responder to assist in the identification, acquisition, and preservation of digital evidence at a crime scene.**

Keywords—mobile device forensics; digital forensics; forensic process, forensic models; MIT App Inventor

I. Introduction

Digital Forensics involves the identification, preservation, collection, examination, and analysis of digital devices. These devices include, but are not limited to, digital cameras, flash drives, computers, internal and external memory drives, mobile devices, etc. Some mobile devices that can be examined include graphic tablets, cell phones, smart phones, CDs, DVDs, and MP3s. Digital evidence has to be collected under certain parameters as to maintain the integrity of the investigation. This process is referred to as a forensic process. While there is not a concrete set of rules for the forensic process there are models that have been proposed to aid in trying to eliminate damage and contamination that can occur at crime scenes.

This paper identifies the types of damage and contamination that can occur at crime scenes when inexperienced first responders arrive at the scene; in addition, we discuss the models that address the preservation and acquisition of evidence at crime scenes, and also explore possible solutions to aid first responders in utilizing techniques to preserve digital evidence at the scene of the crime. In this paper, we propose the development and implementation of a mobile application that first responders can create and use as a guide when identifying, preserving, collecting, and securing evidence. As a result, this application would be useful in assisting first responders during the acquisition process of a digital forensics investigation.

II. Background

In the 21st century, computer crimes have become more of a concern than in past years. The advancement of technology has led to the advancement of crime, such that there is now a need for various methods of evidence collection. Traditionally, physical evidence was collected from a crime scene. Due to the elevation of technology and the rise of digital devices, many of these electronic devices are used in criminal activity. The United States Department of Justice (USDOJ) created a table that categorizes types of crimes and types of evidence associated with those crimes. In Fig. 1, in the sex crime category, where prostitution is being investigated, an investigator should check databases, e-mail, notes, letters, financial/asset records, medical records, address books, calendar, and customer database/records to retrieve evidence [8]. Forensic analysts and investigation teams are responsible for obtaining evidence of this magnitude; however, first responders are often responsible for identifying and collecting devices that this evidence may reside in.

During investigations, first responders are initially deployed to the scene of a crime. First responders, who may or may not be trained forensic examiners, may have dual roles in an investigation. Many times they are untrained in the areas of digital forensic evidence collection and digital crime scene preservation which are vital to any digital forensic investigation. At this point, errors would lead to contamination of evidence and the integrity of the investigation would become compromised or deemed invalid for submission in court proceedings. As the age of computers and technology increase and advance, the crimes committed, where digital devices are involved, will also evolve in type, complexity, and damage perimeter. Thus, the forensic process of digital devices has to be as thorough and concise as possible to protect against viruses, worms, malware, and other possible cyber attacks.

The USDOJ created a forensic process model that guides first responders to help them better assess crime scenes upon initial response. They stated that “the process of collecting, securing, and transporting digital evidence should not change the evidence, digital evidence should be examined by those trained specifically for that purpose, and everything done during seizure, transportation, and storage of digital evidence should be carefully documented, preserved, and available for review” [8]. Their model consists of four phases: collection, examination, analysis, and reporting. First responders are primarily responsible for the collection of evidence at the crime scene. The collection phase is described as the phase in which the search, seizure, and documentation of evidence takes place [1].

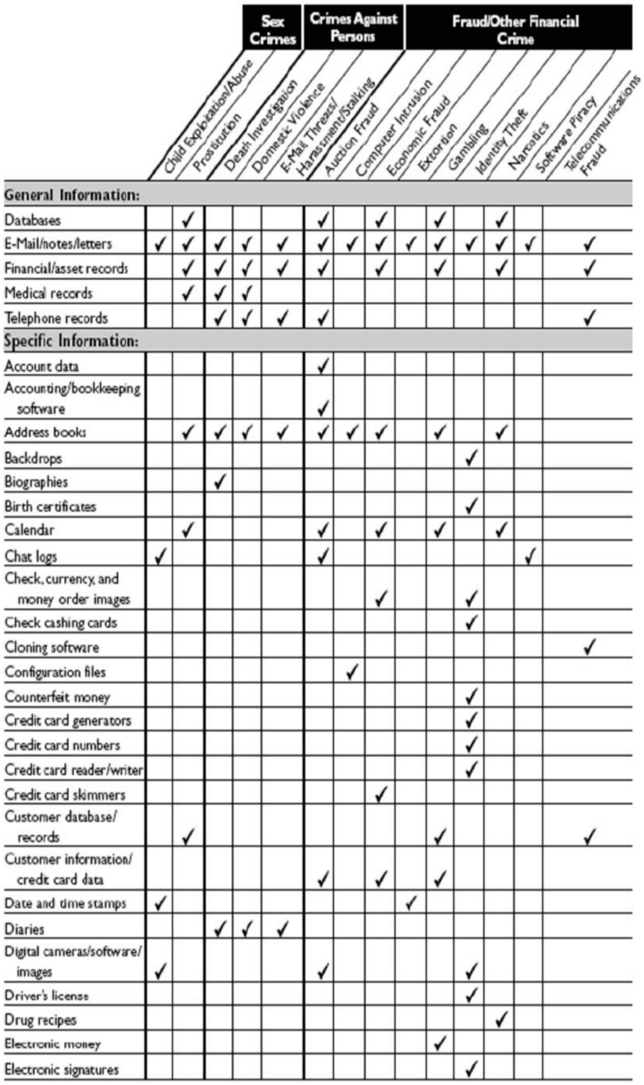
We would like to acknowledge the Department of Energy/National Nuclear Security Administration for providing funding for this research under Dr. Samuel P. Massie Chairs of Excellence Grant#: 240946.

There are a number of other models that have been created and proposed as well [1, 4]. First Responders have to be careful in their seizure of digital evidence because if it is handled improperly it could violate the Electronic Communications Privacy Act of 1986, the Privacy Protection Act of 1980, and federal laws [8]. As a result, researchers have proposed different types of models that provide a more in depth analysis to how a first responder should identify evidence, collect it, and preserve the crime scene until the appropriate forensic teams are deployed to continue the investigation [1, 2, 4, 8].

Fig. 1.1 displays information regarding evidence first responders should collect [8]. These tables created by the USDOJ, categorize the types of evidence one should look for or investigate depending on the type of crime. For example, if it is a sex crime that involves child exploitation/abuse an investigator should investigate e-mail, notes, letters, chat logs, date and time stamps, digital cameras, software, and images [8].

Forensic process models are useful in assisting with breaking down the phases into specific and less ambiguous tasks that will aid in gathering evidence located at any crime scene. The Abstract Digital Forensics Model breaks these two phases into five phases: Identification, Preparation, Approach Strategy, Preservation, and Collection, which first responders are responsible for [1]. These stages merely suggest that first responders identify the type of incident that has occurred, prepare all necessary tools and techniques, and obtain proper authorization to proceed with evidence collection, create a strategy to approach collecting the evidence without tainting it, preserve the state of the evidence whether it is digital or physical, and collect the evidence using proper forensic procedures. These phases require that, during evidence collection, evidence should be authenticated and valid for court proceedings.

Additional models exist that suggest that there is more to be evaluated and there are some models that seem to eliminate the first responder or merely suggest that first responders become trained in forensically sound evidence collection. In most cases, first responders, investigators, and examiners have little or no knowledge of digital forensic process models. Generally, these individuals acquire evidence at the scene based on general evidence collection procedures and/or training from a colleague, in which, the evidence information is documented using paper-based methods. In this paper, we propose the development of a mobile application that can be used as a guide for collecting digital evidence at a crime scene. Using simple, easy-to-learn tools, the application could also be developed by a trained investigator to use to train novice first responders, or it could be developed by the novice first responder himself. This mobile application will serve as a tool to guide first responders, whether trained or untrained, in maintaining the integrity of digital evidence while collecting digital evidence during an investigation. No previous work was found that used the MIT App Inventor software to create an application to assist in the digital forensic investigation process.



	Sex Crimes		Crimes Against Persons			Fraud/Other Financial Crime								
	Child Exploitation/Abuse	Prostitution	Death Investigation	Domestic Violence	E-Mail Threats/ Harassment/Stalking	Auction Fraud	Computer Intrusion	Economic Fraud	Extortion	Gambling	Identity Theft	Narcotics	Software Piracy	Telecommunications Fraud
General Information:														
Databases		✓				✓		✓		✓		✓		
E-Mail/notes/letters	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Financial/asset records		✓	✓	✓	✓	✓		✓		✓		✓		✓
Medical records		✓	✓	✓										
Telephone records			✓	✓	✓	✓								✓
Specific Information:														
Account data						✓								
Accounting/bookkeeping software						✓								
Address books		✓	✓	✓	✓	✓	✓	✓		✓		✓		
Backdrops											✓			
Biographies			✓											
Birth certificates											✓			
Calendar		✓				✓		✓		✓		✓		
Chat logs	✓					✓							✓	
Check, currency, and money order images								✓			✓			
Check cashing cards											✓			
Cloning software														✓
Configuration files							✓							
Counterfeit money											✓			
Credit card generators											✓			
Credit card numbers											✓			
Credit card reader/writer											✓			
Credit card skimmers								✓						
Customer database/ records		✓								✓				✓
Customer information/ credit card data						✓		✓		✓				
Date and time stamps	✓								✓					
Diaries			✓	✓	✓									
Digital cameras/software/ images	✓					✓					✓			
Driver's license											✓			
Drug recipes												✓		
Electronic money										✓				
Electronic signatures											✓			

Fig. 1. Snapshot of USDOJ evidence targets by case category [8]

III. Design and Implementation

The ideas for this application stemmed from a lawyer who suggested the creation of a mobile application that would help professional investigators collect and authenticate evidence using an Android device. According to the lawyer, the Android device should have the dual camera (front and back cameras) capability. Features the application should possess included the following: 1) The ability to snap a photo with outward-facing (back) camera; 2) The ability to launch the user-facing (front) camera to collect video of the first responder at the scene; and 3) To provide scripts for user to say on video to authenticate the attached photos and any evidence collected. Based on the template provided, we modeled a portion of the mobile application's design after the suggested features [6].

The proposed mobile application could be used as a guide for first responders in digital forensic evidence collection. This application would not only provide instructions for evidence identification but could also provide tips and suggestions for evidence collection. The application would have email, web, video, camera, voice and sound features. It would also allow a first responder to record their identity and authenticate their investigation through the photo and video capabilities, access the internet to send secure emails of any of the photos or videos that have been taken, and store any contacts or information that may be needed. This fully functioning application would also have instructions for evidence collection embedded in buttons. This application is an ongoing project and additional modifications are currently in progress.

MIT App Inventor is software originally created by Google Labs for Google engineers and Google users interested in simplified mobile application creation and development [5, 7]. This software is freely available to the public for use. The software is used as a canvas for mobile app design and provides a foundation for inexperienced and experienced programmers. User created MIT App Inventor applications can range from simple games to complex apps that educate and inform. MIT App Inventor's text to speech capabilities allows the phone to ask questions aloud. To use MIT App Inventor, you are not required to be a professional coder. This is because instead of writing code, you visually design the way the application looks and use blocks to specify the application's behavior. MIT App inventor does not require an extensive knowledge in Java programming because the programming aspect is like a puzzle [7]. There is not any necessary hard coding required, the block editor allows the developer to piece the "code" together and then test the functionality of the app through the software's emulator [5, 7].

Figures 1.2 and 1.3 show the components created using MIT App Inventor that provides this application's functionality.

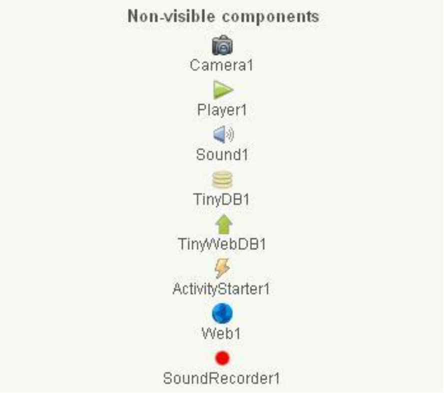


Fig. 2. MIT App Inventor non-visible components [7]

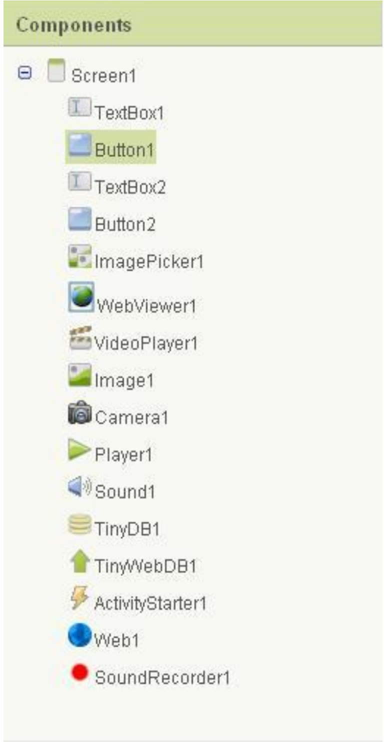


Fig. 2. MIT App Inventor visible components continued [7]

In Fig. 1.2, the non-visible components are not seen on the applications home screen but are still included in the application's functionality. Some of these components are self-explanatory while others are linked together to perform a function. For example, the player1 component allows audio and video to be played and controls the phone's vibration. The tinyDB1 components are storage components that allow the user to retrieve and store information to the phone [7]. These components are helpful to the forensic investigation process because all the information taken can be stored to the device's memory without risk of losing any evidence collected. An application of this magnitude allows a first responder to pause their investigation at any point without losing any of the evidence or data they have obtained. Some applications have to have web access to operate. This application would not require web access to function but would have web capabilities in case a first responder needed to find the type of model of a digital device left at the scene of a crime. This application is designed to operate on Android mobile devices that have dual camera capabilities; however, at the time of development, the MIT App Inventor software did not provide that component. If an Android mobile device has a camera, the application can still run. While they may not be able to have dual camera functionalities, a sound recorder has been implemented in the application's design to allow the first responder to identify themselves and record any information necessary to the investigation. This tool is not designed to be restricted to cell phone usage only. Graphic tablets such as the Samsung Galaxy, Toshiba Thrive, and other devices that use the Android OS software can access it.

Fig. 1.3 depicts the app inventor emulator with some of the mobile application functions provided on the home screen. The sound recorder would also list the tasks to be completed based on the model phases previously presented. Embedded in the picture with the magnifying glass is a recorded sound that instructs a first responder as to the information they need to provide to authenticate their identity. We discovered in development that a sound recorder would need to be added as well to accommodate those devices not equipped with front and back (dual) camera capabilities. Therefore, a sound recorder and a note pad were added to the apps functions. Google app inventor labs allows mobile app development to be created through the blocks editor, which allows inexperienced or beginner programmers to experiment with application development and a simpler type of coding.

The Blocks Editor component of the app inventor software is used for visually programming the application by fitting the pieces of those blocks together sort of like a coding puzzle as shown in Fig. 1.4 [7]. The software itself will let you know if you are coding incorrectly. Additionally, the software is equipped with programming capabilities for those who have experience in Java programming. Future goals are to enhance this application by providing the user with the ability to link to web pages that have information about evidence collection process on them and explain how to identify the types of evidence for the type of crime. Resultantly, this application could serve as a guide for first responders and also an evidence tool in the law enforcement community.

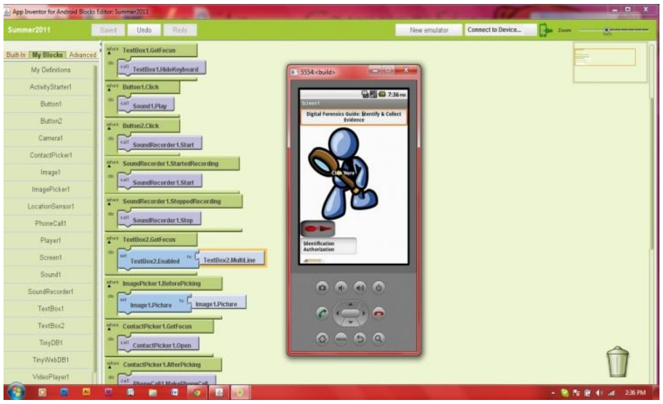


Fig. 3. App Inventor Block Editor and Emulator for the app

IV. Testing Phase

Testing the tool involved numerous trials resulting in some improvements in the applications development as well as some limits in the app inventor software. At the time of testing, some of the application's capabilities are not supported by the app inventor software and does not transfer very well to an actual cellular device. Some hard coding using logic and Java or JavaScript could possibly solve this problem however; extensive work on the app is required. During the testing phases, we discovered that the emulator accurately showed how the application functioned on a mobile device. At the time of development, some of the bugs in the web component were currently undergoing construction, and the software did not provide for the notebook capability which allowed attaching comments to a photograph. These are some of the limitations that were presented during testing phases.

Although problems were encountered in the application, some of the original components did function properly. The figures below depict the applications capabilities thus far.

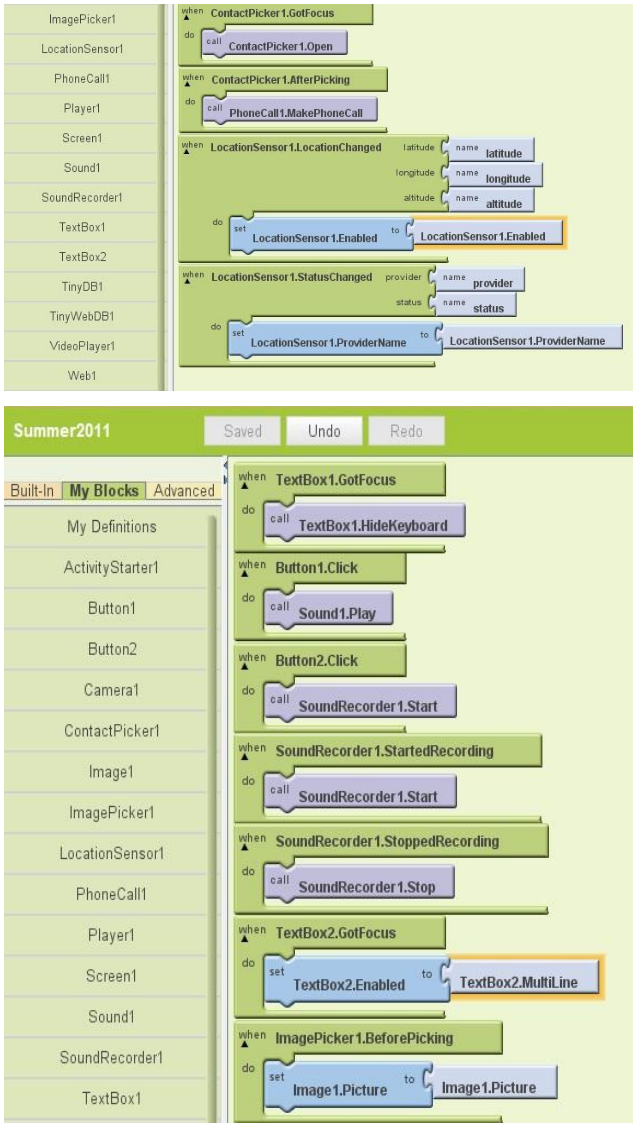

ImagePicker1
LocationSensor1
PhoneCall1
Player1
Screen1
Sound1
SoundRecorder1
TextBox1
TextBox2
TinyDB1
TinyWebDB1
VideoPlayer1
Web1
when ContactPicker1.GotFocus
do call ContactPicker1.Open
when ContactPicker1.AfterPicking
do call PhoneCall1.MakePhoneCall
when LocationSensor1.LocationChanged
latitude name latitude
longitude name longitude
altitude name altitude
do set LocationSensor1.Enabled to LocationSensor1.Enabled
when LocationSensor1.StatusChanged
provider name provider
status name status
do set LocationSensor1.ProviderName to LocationSensor1.ProviderName
Summer2011
Saved
Undo
Redo
Built-In
My Blocks
Advanced
My Definitions
ActivityStarter1
Button1
Button2
Camera1
ContactPicker1
Image1
ImagePicker1
LocationSensor1
PhoneCall1
Player1
Screen1
Sound1
SoundRecorder1
TextBox1
when TextBox1.GotFocus
do call TextBox1.HideKeyboard
when Button1.Click
do call Sound1.Play
when Button2.Click
do call SoundRecorder1.Start
when SoundRecorder1.StartedRecording
do call SoundRecorder1.Start
when SoundRecorder1.StoppedRecording
do call SoundRecorder1.Stop
when TextBox2.GotFocus
do set TextBox2.Enabled to TextBox2.MultiLine
when ImagePicker1.BeforePicking
do set Image1.Picture to Image1.Picture

Fig. 4. MIT App Inventor Blocks Editor

Fig. 1.5 depicts the emulator in the app inventor software using the multi-line note pad capability of the app. During development and testing, it was found that this feature did not allow the user to connect comments to actual digital media (pictures, video, etc). Comments on pictures were not permitted by this application at this stage of development. Ideally, the user would be able to type comments associated with the picture that they have selected to view. Attempts to integrate this feature are ongoing at this time. In our continuous development efforts, we plan to address and correct this issue.

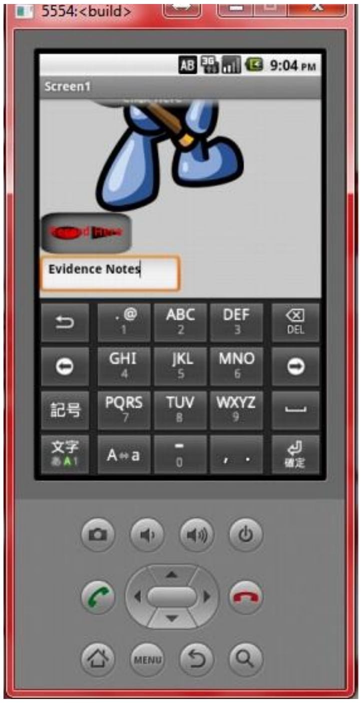
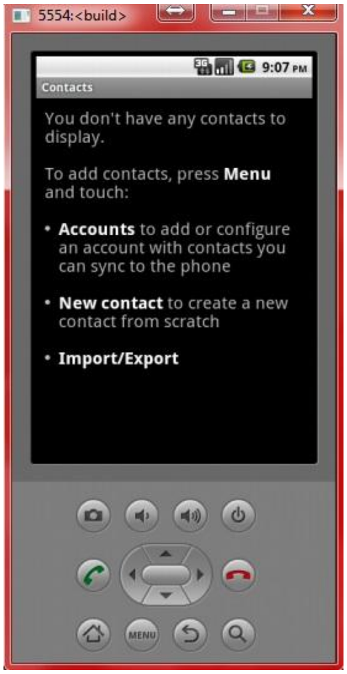


Fig. 5. App Inventor Emulator displaying the entering of evidence notes



You don't have any contacts to display.

To add contacts, press **Menu** and touch:

- **Accounts** to add or configure an account with contacts you can sync to the phone
- **New contact** to create a new contact from scratch
- **Import/Export**

Fig. 6. MIT App Inventor Emulator displaying stored contacts

Fig. 1.6 shows the contacts feature of the application working properly. If the user clicks the contacts button it will link to whatever contacts are stored of the mobile device. The emulator did not show any contacts in the figure because there were not any stored on the device. However, it did perform correctly with the contacts when they were added.

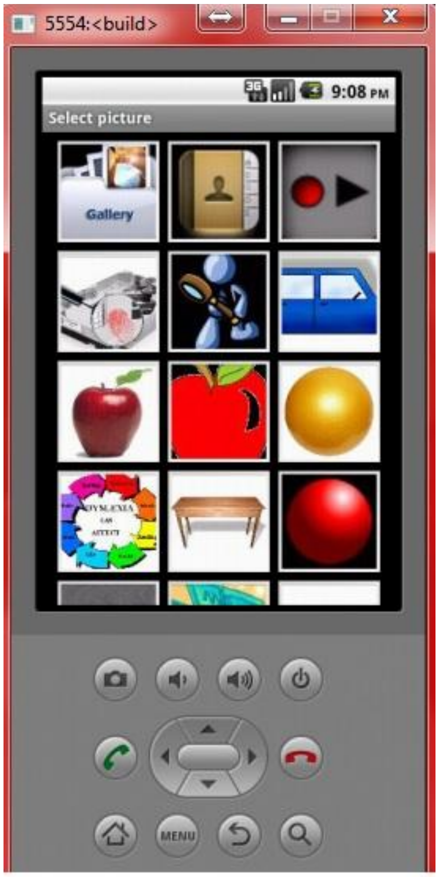


Fig. 7. MIT App Inventor Emulator displaying app selections

Fig. 1.7 depicts the photo gallery feature of the mobile application as a functioning component. The figure depicts all the images related to this mobile app and other pictures downloaded and stored in a folder pertaining to mobile application design. While the emulator only shows those images in the folder, testing showed that the gallery function linked to the mobile phone's photo gallery and allowed pictures to be stored.

V. Conclusions

In this paper, we discussed the importance of documenting digital evidence at the crime scene, we identified the different types of incidents that could occur when first responders arrive at the scene, and we discussed why a mobile application would be useful in the evidence identification and collection of first responders. Little or no research has been found that discussed how mobile applications could be used to assist first responders in the evidence identification, collection, and documentation process. Therefore, in this paper, we attempted to develop a mobile application that could assist first responders in their digital investigations. However, the tool is currently undergoing modifications. Furthermore, it could still be developed into a powerful training tool that could be used by law enforcement and other investigative teams during an investigation.

During the development of the digital forensics application, Google App Inventor was used. MIT recently began hosting Google's App Inventor, which is now MIT App

Inventor in March 2012 [7]. Most of the design and experimentation performed prior to March 2012, shows that some of the components of the initial design were not supported by the Google App Inventor Emulator or software at the time of development. For example, we embedded a web viewer and email function in the application's initial design but testing has shown us that these functions are not supported by Google App Inventor's Emulator. This mobile application can be extended to include modifications, additional functionalities, and testing with real investigators. It could provide mobility and a succinct, electronic way of storing and documenting evidence acquired at the crime scene. It could also be modified to implement each of the phases of the digital forensic process to aid investigators during the analysis, reporting, and presentation phases. Given the simplicity of the tool, investigators could develop their own custom applications, for no cost to them, based on their individual needs. This application could be used as a basis for creating other versions of this application as well. The applications could be modified to include games, which provide different scenarios/crime scenes, for novice first responders to enhance their evidence identification and recovery skills.

Given the increase in digital crimes and the need for skilled digital forensic investigators, the development of such a tool is needed. Law enforcement officers generally do not have the time or funds to engage in training, especially with training on new digital tools. Developing a mobile application that can assist first responders in maintaining the integrity of the evidence and documenting the evidence "in real-time," benefits not only the first responders, but the entire law enforcement community.

VI. FUTURE WORK

Future enhancements to this work include linking a notepad and email function to the photo gallery to allow comments to be written associated with the photos and securely encrypted and emailed and stored on that organization's servers. Also, the bugs in the web viewer could be deciphered to allow email functionality and web browsing capabilities. An additional function to be added to the application is the implementation of a sound recorder that could be embedded in the application or link to a phone that allows voice recording to support those Android phones that are not dual camera capable. These additional enhancements to the application would make it both a useful and beneficial digital forensic investigation tool.

REFERENCES

[1] A. Tanner and D. Dampier, "An approach for managing knowledge in digital forensic investigations," International Journal of Computer Science and Security, vol. 4, no. 5, pp. 451-465, December 2010.

[2] F.C. Dancer, D. Dampier, J. Jackson, N. Meghanathan, "A theoretical process model for smartphones," Proceedings of the Second International Conference on Advances in Computing and Information Technology (ACITY), July 2012.

[3] R. P. Milsan, "Cell phone Crime Solvers: Could the murder victim's Blackberry lead to her killer?" IEEE Spectrum, July 2010.

[4] A. Ramabhadran, Security Group, and T. Elxsi, "Forensic investigation process model for Windows mobile devices," http://www.forensicfocus.com/downloads/windows-mobile-forensic-process-model.pdf, 2012.

[5] Google Labs, "App Inventor for Android," https://code.google.com/p/app-inventor-for-android/, 2011.

[6] App Inventor Coffee Shop, "Inspiration for a useful, even marketable,

[7] app," 2010.

[8] MIT App Inventor, "Explore MIT App Inventor," http://appinventor.mit.edu/explore/, 2012.

[9] A. Tanner, " A concept mapping case domain modeling approach for digital forensic investigations," doctoral thesis, Department of Computer Science and Engineering, Mississippi State University, Mississippi State, MS, 2010.

Development of Social Media GIS for Information Exchange between Regions

Syuji YAMADA
Graduate School Student,
Graduate School of Information Systems,
University of Electro-Communications
Tokyo, Japan

Kayoko YAMAMOTO
Associate Professor,
Graduate School of Information Systems,
University of Electro-Communications
Tokyo, Japan

Abstract—**This study aims to develop a social media GIS (Geographic Information Systems) specially tailored to information exchange between regions. The conclusions of this study are summarized in the following three points.(1) Social media GIS is a geographic information system which integratesWeb-GIS, SNS and Twitter into a single system. A social media GIS was conducted for the collection of regional information in the eastern part of Yamanashi Prefecture. The social media GIS uses a design which displays its usefulness in multi-directional information transmission and in easing the limitations of space, time and continuity, making it possible to redesign systems in accordance with target cases.(2) During the operation of the social media GIS for about two months, most of the users were in their20s. Users exchanged regional information using the comment and button functions.(3)The system was evaluated based on the results of a questionnaire to users and an access analysis oflog data during operationin order to identify measures for improvement of the system. Because of users' high evaluations of its original functions, the overall operability of the system was highly evaluated. Most of the contributed information was only known to local residents, and it was evident that the system fulfilled its intended role.**

Keywords—Information Exchange;Web-GIS; Social Media; SNS; Twitter

I. INTRODUCTION

In recent years in Japan, where the formation of a highly information-oriented society is being achieved, the amount of information about urban areas is on the increase, and a diverse range of information can easily be obtained by various means anywhere, anytime.However, in regions outside urban areas, although the amount of information is increasing, compared with urban areas, it can by no means be termed sufficient.Further, it is difficult for people other than those who reside, commute to work, or attend schools in the regions outside urban areas to obtain and utilize detailed regional information.For example, considering the case of tourism in regions outside urban areas, local governments, tourist associations and other organizations use websites to transmit regional information, but this information alone cannot be termed sufficient in either amount or content.Further, other than for famous tourist spots, not much detailed regional information is published in guidebooks or on websites related to tourist information.Therefore, when tourists visit a region for the first time, they often feel inconvenience in regard to obtaining regional information.

Accordingly, in regions outside urban areas, it is necessary to focus on the importance of "regional knowledge", which is information, knowledge and wisdom created when highly specialized "expert knowledge", based on scientific knowledge, and "experience-based knowledge", produced by the experiences of local residents in those regions, are combined, and to utilize this regional knowledge (Science Council of Japan, 2008)[1].However, concerning experience-based knowledge, which is the part of regional knowledge that is possessed by local residents, in many cases the knowledge is possessed by individuals and not communicated to others.Therefore, it is essential to change experience-based knowledge, which exists as "implicit knowledge", remaining untold to others and not visualized, to "explicit knowledge", which is knowledge in a form that can be accumulated, arranged, utilized, and made available to the public.Further, it is essential for the knowledge to be accumulated and shared among people in the region, and for information to be exchanged with people in other regions.

Meanwhile, in modern Japan, the transition from narrowband (ISDN) to broadband (ADSL, FTTH) has been made, and a stable internet environment has been rapidly provided.Further,Web-GIS, which enables more effective use of geographic information systems (GIS) information provision and sharing functions, is starting to attract attention.In order to provide GIS as systems which many people can use over the long term, combinations withWeb-GIS, which can visualize the actual region on the website and enable editing of information if necessary, and systems which include the distinctive functions of SNS, which allow target users to be narrowed down in advance, can be proposed.Further, the aim of this study is not just accumulation and sharing of regional information, but also exchange of information between regions; therefore, the development of a social media GIS which enables operation together with other social media in addition toWeb-GIS and SNS is necessary.Taking these background factors into account, this study aims to develop a social media GIS which enables accumulation and sharing of regional information and exchange of information between regions, in order to supplement the scarcity of information in regions outside urban areas.

This study was conducted according to the following framework and method. First, the preceding studiesin the related fields and the originality of this study were introduced (Section II). Next, the design (Section III) and development (Section IV) of a social media GIS especially tailored for the aim of this study were independently conducted. Moreover, two patterns of users aged 18 years and over - those inside and those outside the region targeted for operation – were anticipated, and operation testing and operation of the social media GIS (Section V), as well as evaluation of the system and identification of measures for improvement (Section VI) were conducted. Here, it was anticipated that each user would use the system for approximately one month, and the operation test and evaluation of the operation test were conducted; following that, actual operation was conducted. Further, web questionnaires targeting the users and access analysis using log data were conducted, and based on the results of those, evaluation of the system was conducted, and measures for improvement were identified.

II. Related Works

Preceding studies concerning regional information which employs GIS and web applications that are in fields related to this study can be broadly divided into three groups: (1) Studies concerning accumulation and sharing of regional information which employ information systems; (2) Studies concerning the design and development of Web-GIS; and (3) Studies concerning regional information contributed to microblogs.

In group (1), Itou et al.(2005)[2], Tsuboi et al.(2008)[3], Nuojua et al.(2008)[4], Shimizu et al.(2008)[5] and Yu et al.(2009)[6] demonstrate the possibility of accumulating and sharing information using GIS for the purposes of regional stimulation and public participation. Umeda et al.(2006)[7], Hara et al.(2008)[8],Kitahara et al. (2009)[9] and Chow et al.(2010)[10] share information using SNS.In group (2), Kirimura et al.(2008)[11], Soga et al.(2008)[12], Nuojua(2010)[13], Hosoya et al.(2011a, 2011b)[14,15]and Kubota et al.(2012)[16] demonstrate the necessity of system design which takes into account convenience, usefulness and operability, and the necessity of effective use ofWeb-GIS suited to their intended themes.Further, Yanagisawa et al.(2012)[17] aims at accumulating and sharing regional information, and Nakahara et al. (2012)[18] aims at supporting communication between users.In these studies, systems which integrateSNS,Web-GISand Wiki were designed and developed, the systems were actually operated in regional communities, and the operations were evaluated.In group (3), through acquisition and analysis of information contributed to Twitter, Fujisaka et al. (2010)[19]and Lee(2012)[20] estimate the extent of the influence of regional events, Fueda et al. (2012)[21] investigate tourist information, and Sagawa et al.(2012)[22]propose a website which assists in providing information about daily life.Further, Lee et al.(2010)[23],Lee et al.(2011)[24] and Hashimoto et al. (2012)[25]detect geo-tagged tweets, and Cheng et al. (2010)[26] and Hiruta et al.(2013)[27] detect “location-triggered” geo-tagged tweets.

Further, in addition to what has been conducted in studies, some regional SNS accumulate and share information using digital maps.A representative example is that the websites “HYOCOM” of Hyogo Prefecture and “Hamatch!” of Yokohama City in Kanagawa Prefecture started internet comment maps which employ Google Maps in 2006 and 2007 respectively, and are accumulating and sharing information concerning recommended shops and places that is contributed by local residents.Further, in 2003, the website “gorottoyacchiro” of Yatsushiro City in Kumamoto Prefecture also began accumulating and sharing regional information and information about daily life contributed by local residents on a digital map.However, in these examples from regional SNS, the focus is information contribution and browsing functions, and there are few functions for users, such as functions for photo contribution and publication, and for information exchange between users.Further, there is a problem relating to interface, in that the digital map on the screen of the user is small, so it is difficult to read map information.

In groups (1) and (2) of the above-mentioned preceding studies, the aim of the studies is accumulation and sharing of information, and communication between users.Web-GIS alone or systems developed by integrating web applications such asWeb-GIS, SNS and Wiki into a single system are used.In group (3), study in which Twitter, which is being used as a method for acquiring regional information, is integrated withWeb-GIS or other web applications has not been carried out up to the present.Therefore, compared with the results of the preceding studies listed above, this study is unique in that it accomplishes both relief of user stress and accumulation of information, which tends to be given no importance in microblogs, by enabling the following: the ease of contributing information using microblogs; accumulation and sharing of information within communities and exchange of information between regions, conducted using a self-developed SNS; and function and interface design which takes into account problem areas related to examples of the use of digital maps in regional SNS.Further, the value of this study lies in the fact that after the system was developed, an interview survey which targeted subjects of an operation test was conducted, and detailed system configuration which anticipated the everyday use of local residents was conducted; and following that, in order to verify the effectiveness of the system, the system was actually operated and evaluated in the region for operation.

III. Design of the system

A. Characteristics of the system

In this system, as shown in Fig. 1, three web applications, that is, aWeb-GIS, a SNS and Twitter, were integrated to develop a social media GIS that is effective for information exchange between regions which is based on the accumulation and sharing of regional information.The method for integrating these three web applications was to include theWeb-GIS in the SNS, and conduct a mashup using the SNS and Twitter.The system enables geographical understanding of location information relating to information contributed, via theWeb-GIS; management and visualization of information contributed on the digital map which includes environment variables; accumulation and sharing of regional information of users and exchange of information between regions using the self-developed SNS; and classification of the importance of contributed information.Further, by enabling the contribution

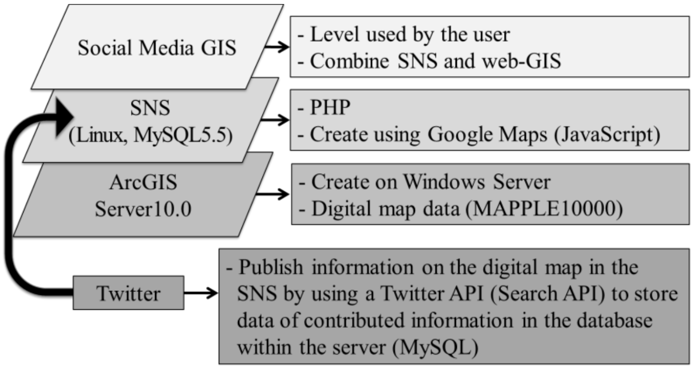

Social Media GIS
- Level used by the user
- Combine SNS and web-GIS
SNS
(Linux, MySQL5.5)
- PHP
- Create using Google Maps (JavaScript)
ArcGIS
Server10.0
- Create on Windows Server
- Digital map data (MAPPLE10000)
Twitter
- Publish information on the digital map in the SNS by using a Twitter API (Search API) to store data of contributed information in the database within the server (MySQL)

Fig. 1. System design

of information from Twitter as well, user stress is relieved and long-term operation is realized; further, users inside and outside the region of operation can use Twitter to light-heartedlycontribute information from a portable information terminal anytime, regardless of whether they are indoors or outdoors.Based on the above, the usefulness of the system, which was outlined in Section II, is described in detail below.

1) Interactivity of information transmission

Many local government websites transmit various types of information relating to administrative services and the region.However, in many cases, one-sided transmission of information from the administrators is the prevailing pattern, and there are few situations where the voices of local residents are reflected.Therefore, in this system, all users possess an account, and thereby each user has responsibility for information transmission, and interactive information transmission is enabled.

Further, because information transmission is conducted using digital maps as a base, knowledge and wisdom possessed only by local residents of the region for operation can be accumulated and shared as contributed information, with location information attached, in addition to photos.Based on this, interactive communication between users inside and outside the operation region can be conducted; therefore, exchange of information between regions is made possible.Accordingly, it can be anticipated that even when users are not residents of the region for operation, acquiring regional information published on the digital map will enable them to make appropriate selections concerning activities when they visit the region for operation.Further, it can be anticipated that accumulation and sharing of regional information between all users will be promoted.

2) Relief of spatial and temporal limitations

As a situation in which these limitations may arise, a situation in which a user is outside and cannot connect to the internet from a computer may be imagined.In order to ease such restrictions, a mashup was performed with a web service which enables use from a portable information terminal as well, and therefore, information contributions can be made from portable information terminals, and the system can be used anytime, regardless of whether the user is indoors or outdoors.In particular, thanks to the mashup with Twitter, in cases where users have made a new discovery or have information they wish to share with other users, because location information is automatically acquired from a GPS via the portable information terminal, simply by tweeting on Twitter, users can contribute regional information to the system independent of spatial and temporal limitations, without having to check location information themselves.

3) Ease of restrictions concerning continuing operation

In order to accomplish long-term maintenance of an environment in which interactive transmission of information unlimited by place and time is enabled by carrying out the steps outlined above in parts (1) and (2), it is necessary to design a system which can manage information which gathers piece by piece.Further, in the case where an open platform is employed, if a structure which can manage contributed information thoroughly does not exist, when inappropriate posts or items of information based on biased evaluations are contributed, it becomes difficult to conduct operation suited to the purpose of the system.Therefore, in this system, continuous operation is enabled by conducting centralized control of contributed information using MySQL, and checking and identifying inappropriate users via account management which employs the SNS.

B. System operating environment

In the design and development of this system, a leading method for web system building called XAMPP was used.Further, in this study, because the burden of processing in theWeb-GIS is very large, the web server and the GIS server were set up separately and split up, in order to speed up processing.Fedora Core 3 from the Fedora Project was used for the web server operating system.The SNS used in this system was developed independently using JavaScript and PHP, and for the database, MySQL 5.5 was used.For the development of the server for theWeb-GIS, Microsoft Corporation's Windows Server 2008 and Esri, Inc.'s ArcGIS Server 10.0 were used.For theWeb-GIS, there are two main types of users; that is, computer users and portable information terminal users.For each type of user, JavaScript receives an action from a user, and following that, it starts applications.In the browser compatibility test, the Spoon web service was used.

Further, recent portable information terminals are shifting from a style such as that of traditional portable phones, in which websites developed especially for portable terminals are browsed, to a style such as that of smart phones, in which websites edited especially for computers can also be browsed.When operating a system which handles digital maps, if the portable information terminal used has a touch panel, it is easier to perform pointing operations on the map than it is when limited operations such as those of arrow keys are used.Further, the larger the screen on which the digital map is depicted, the easier it is to view the digital map.In addition, Impress R&D (2012)[28] pointed out that about 40% of internet users aged thirteen years and over were smart phone users in 2012 in Japan, and if that trend continues, it is possible that in 2014 the majority will be smart phone users.Impress R&D also indicated that smart phone use is becoming widespread among people of middle age and advanced age as well.Taking the above-mentioned factors into consideration, this study focuses on both conventional mobile phones and smart phones as portable information terminals.

C. Outline of system design

As described in detail in this section, concerning the three types of web applications, by conducting independent design specialized for the purpose of this study, which is exchange of information between regions based on accumulation and sharing of regional information, a social media GIS can be suitably operated in the region intended for operation.

1) Web-GIS

In this study, for theWeb-GIS, Esri, Inc.'s ArcGIS Server 10.0 was used, and for theWeb-GIS digital map data, the SHAPE version (Rel.8) of Shobunsha Publications, Inc.'s MAPPLE10000, which is part of their MAPPLE digital map data and includes detailed road system data, was used.As the map that was combined with this map data, the user interface(UI) of Google Maps was used.Among the options provided by Esri, Inc. that are ArcGIS Server 10.0 API targets, the Google Maps UI was the one used the most in preceding studies in fields related to this study.Concerning the combination of MAPPLE10000 (SHAPE version) and Google Maps, Google Maps uses the new geodetic system coordinates, while MAPPLE10000 conforms to the former geodetic system coordinates; therefore, ArcTKY2JGD, which is provided by Esri, Inc. as product support, was used to convert the MAPPLE10000 geodetic system coordinates to the new ones.Furthermore, editing was performed such that information peculiar to each location could be added using ArcMap 10.0.By using theWeb-GIS, users can refer to detailed road systems which also include minor roads output from MAPPLE10000, and thereby can accurately check the location related to information contributed.

2) SNS

In this system, it was desirable that exchange of information between regions should be conducted voluntarily by many users; therefore, community features were not provided.Further, user personal data registration/profile publication, contribution and browsing of information and photographs, methods for exchanging information between regions, functions relating to classifying the importance of contributed information and so forth were designed independently to suit the aim of this study.

Concerning a method for exchanging information between regions, the system was designed such that text information and a digital map were within the same page, in order to enable efficient digital map-based exchange of information between users.Further, in this study, when there has been some sort of exchange between users inside and outside the region for operation, an exchange of information between regions is said to have been conducted.Specifically, in the case where a user outside the region for operation posts a comment in response to information that a user inside the region for operation has contributed, and the case where a button for classifying the importance of contributed information is clicked, it is considered that an exchange of information has been conducted.Further, the number of times that users inside and outside the region for operation click the button for classifying the importance of contributed information is converted into points, and used in the points ranking of contributed information.

3) Twitter

In this study, in order to realize long-term operation by preventing a decrease in the number of active users, and to design the system such that users inside and outside the region for operation could light-heartedly contribute information anytime using a portable information terminal, regardless of whether they were indoors or outdoors, Twitter was selected from among the different types of social media, and a mashup was performed with a self-developed SNS.Of the various types of social media, Twitter allows the easiest contribution of information, and many tweets per day can be anticipated; therefore, Twitter is indispensable to long-term operation.Further, as mentioned in Section III-B, in this study, taking into consideration the present situation, which is that smart phones are moving out of their introductory period and into their transition period, a system which can be used from both conventional portable phones and smart phones was developed.In order to do this, rather than separately developing one information contribution system for conventional portable phones and another for smart phones, the system was designed such that a mashup with the self-made SNS and Twitter was performed, so Twitter could be used from both the two types of portable information terminals.

IV. System development

A. System front end

In this study, multiple functions have been designed independently as described below, and users can select methods suited to their preferences for each situation; therefore, it can be anticipated that accumulation and sharing of information and exchange of information between regions will be stimulated.

1) Functions for users

a) Personal data registration/profile publication functions

The first time a user makes access to the system, they use the initial registration screen to register personal data such as their "User ID", "Password", "Age group", "Sex", "Region" and "Greeting".This is because it is desirable that the system should be designed such that when users conduct interactive communication with each other, they can be identified to a certain extent.Further, because users who do not wish to make their personal data public have been taken into consideration, the "User ID" is designed such that the user's real name and account name are not specified, and the user can freely select and enter a user ID.Further, when a user logs in after performing the initial registration, they can perform operations on the browsing/information contribution screens and so forth.

b) Information contribution/browsing functions

Two types of methods were provided for when a user contributes regional information – the method of contributing information from a computer, and the method of contributing information from a portable information terminal using Twitter.In the former method, first, the user clicks "Post" on the home page of the website on their computer screen, to go to the posting page.On the posting page there is a form in which the user can enter the "Title" and "Main text".After the user has entered the content into the form, location information relating to the posted content can be added simply by clicking the posting location on the digital map.The "location information" is entered into MySQL, and when transmission is performed, posting is complete.In the latter method, data of information posted using Twitter from a portable information terminal is acquired, and displayed on the digital map on the posting page of the user which is set up in the system.In both methods, at the time of posting, if necessary, a photo image file can also be attached.All contributed information is shared within the SNS.From the home page, the following three sections can be browsed: The ten most recently contributed items of information, a list of contributed information, and the points ranking of contributed information.

Further, contributed information can also be viewed by clicking "View" on the home page.On the viewing page, information shared on the digital map is displayed as a marker.When the marker is clicked, a balloon is displayed, and a more detailed page is moved to.Further, as shown in Fig. 2, a comment function and a mention-later button function can be used.

c) Button functions/ranking function

Button functions are used for classifying the importance of contributed information. Two types of buttons were provided - "I didn't know" for users within the region of operation, and "I want to go there" for users outside the region of operation. Thus, the provision of button functions in this system enables users to easily express their intention in regard to information they have viewed. In this study, when a user outside the region of operation uses the above-mentioned comment function and the "I want to go there" button function in response to information posted by a user in the region of operation, it is defined as an exchange of information between regions. Further, for each item of contributed information, one point is added each time users inside and outside the region of operation click either of the two above-mentioned buttons, and each piece of contributed information is evaluated. Moreover, by including a ranking function which displays contributed information in descending order of total points gained, the system avoids losing contributed information which users are strongly interested in amongst other information.

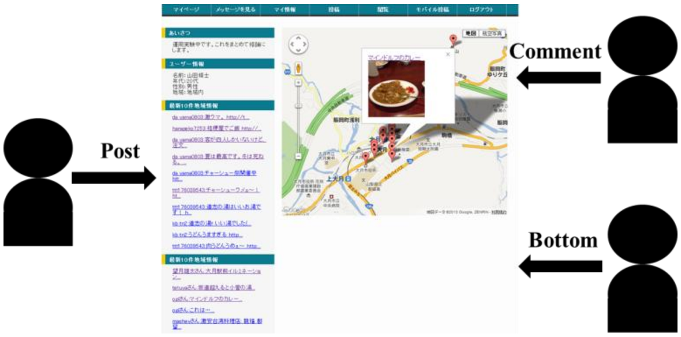


Fig. 2. Exchange of information between regions which employs functions for users

2) Functions for administrators

Administrators log in from a login page exclusively for administrators, and a screen exclusively for administrators is provided.Using the administrator screen, administrators manage users, and in cases where there has been an inappropriate statement or inappropriate behavior, they manage the matter by taking action, such as closing user accounts.Further, administrators can view all contributed information, contributor names, and dates and times of contributions on a screen which lists them; therefore, if by any chance an appropriate posting is made, they can delete it with just one click. Thanks to these aspects of the system, the burden of administrators can be reduced, because there is no need for them to search to check whether there are inappropriate items of contributed information in the system. Further, the case where local residents actually perform the role of administrators in the regional community is anticipated. The system is designed such that MySQL is managed using graphical user interface (GUI) and administrators who do not have a very high level of IT literacy can also manage and administer; therefore, the burden on administrators can be reduced as much as possible.

B. System back end

1) Management system for contributed information that is run by administrators

Information, photos image files, comments and so forth contributed by all the users are all stored in the database of the system as data.Further, administrators can view these items of contributed information on a screen which lists them, so if by any chance an inappropriate contribution is made, it can easily be deleted.

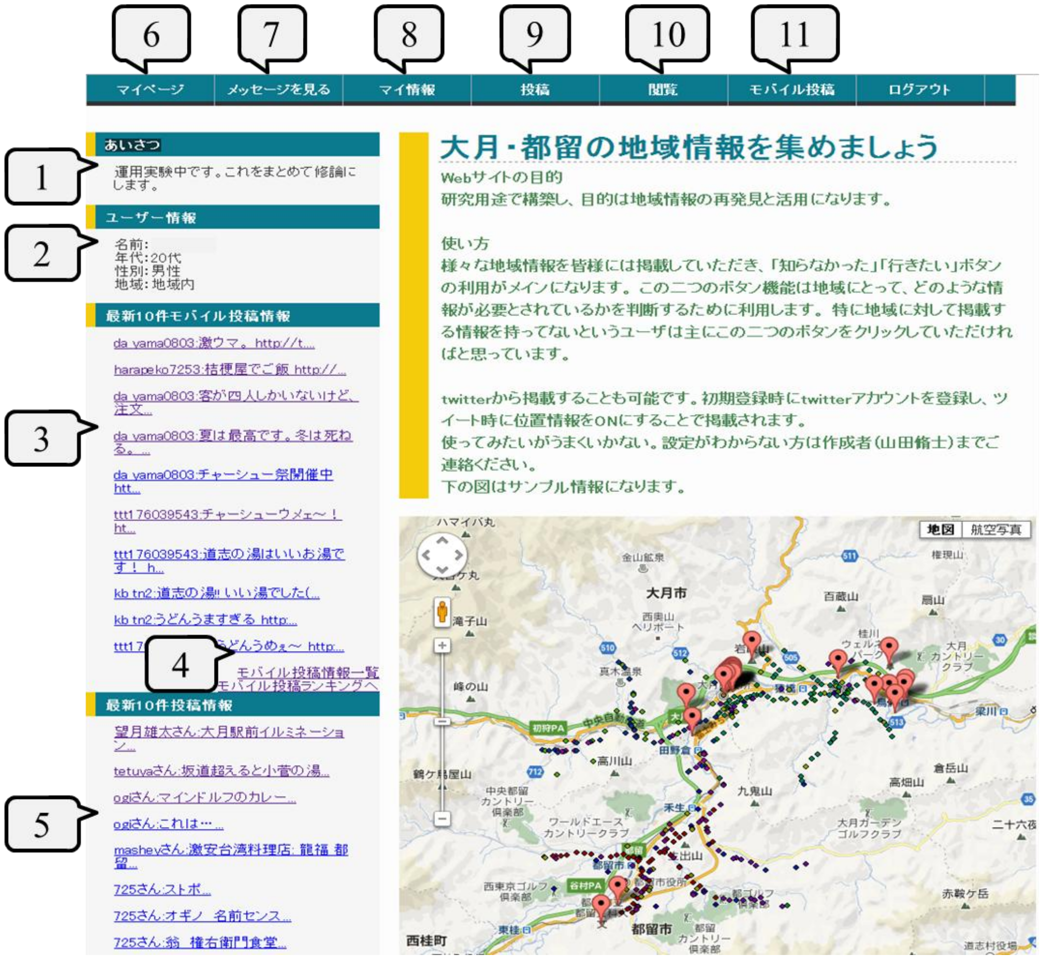
2) Mashup system with Twitter

In this study, when a mashup is performed with Twitter, the Search API with Basic authentication protocol is used, and thereby the effort involved in information contribution is minimized and user stress is reduced.Conventionally, when a Twitter mashup system is developed, the OAuth authentication protocol is often used.However, in this study, upload from the main part of the system to Twitter of information for contribution and so forth is not conducted; therefore, the Search API with Basic authentication protocol, which allows acquisition by searching Twitter data, was employed.

The data for reflection in the system (main text, location information, account names, dates and times) is obtained by making a query specification.In the process for acquiring data of information contributed using Twitter from a portable information terminal, users simply register a Twitter account name in the blank at the time of initial registration.The rest is performed by the back end.Data of information contributed using Twitter of registered users is obtained by applying all account names saved in the database to the "user" portion of "from:<user>" of tags.

C. System interface

The system has three types of interface – the computer screen of the user (Fig. 3), the portable information terminal screen of user, and the computer screen of the administrator.Using the administrator screen, inappropriate contributed



あいさつ

運用実験中です。これをまとめて修論にします。

ユーザー情報

名前:
年代:20代
性別:男性
地域:地域内

最新10件モバイル投稿情報

da yama0803:激ウマ。http://t...
harapeko7253:桔梗屋でご飯 http://...
da yama0803:客が四人しかいないけど、注文...
da yama0803:夏は最高です。冬は死ねる。...
da yama0803:チャーシュー祭開催中 htt...
ttt176039543:チャーシューウメェ～！ ht...
ttt176039543:道志の湯はいいお湯です！ h...
kb tn2:道志の湯!! いい湯でした(...
kb tn2:うどんうますぎる http:...
ttt17...うどんうめぇ～ http:...

モバイル投稿情報一覧
モバイル投稿ランキングへ

最新10件投稿情報

望月雄太さん:大月駅前イルミネーション...
tetuyaさん:坂道超えると小菅の湯...
ogiさん:マインドルフのカレー...
ogiさん:これは…...
masheyさん:激安台湾料理店: 龍福 都留...
725さん:ストボ...
725さん:オギノ　名前センス...
725さん:翁　権右衛門食堂...

大月・都留の地域情報を集めましょう

Webサイトの目的
研究用途で構築し、目的は地域情報の再発見と活用になります。

使い方
様々な地域情報を皆様には掲載していただき、「知らなかった」「行きたい」ボタンの利用がメインになります。この二つのボタン機能は地域にとって、どのような情報が必要とされているかを判断するために利用します。特に地域に対して掲載する情報を持ってないというユーザは主にこの二つのボタンをクリックしていただければと思っています。

twitterから掲載することも可能です。初期登録時にtwitterアカウントを登録し、ツイート時に位置情報をONにすることで掲載されます。
使ってみたいがうまくいかない。設定がわからない方は作成者（山田脩士）までご連絡ください。
下の図はサンプル情報になります。

No.	Description
1	User greeting
2	User profile publication
3	The ten most recent items of information contributed from portable information terminals using Twitter
4	Go to a list of information contributed from portable information terminals using Twitter and the ranking page
5	The ten most recent items of information contributed from computers
6	Go to the home page of the user (sample information is displayed on a digital map)
7	Go to the page which contains messages from administrators
8	Go to the page where change and registration of personal data can be made
9	Go to the page where information can be contributed from a computer
10	Go to the page where information contributed from computers can be viewed
11	Go to the page where information contributed from portable information terminals using Twitter can be viewed

Fig. 3. Illustration of user computer screen and functions

TABLE I. OPERATION PROCESS FOR THE SYSTEM

Process	Aim	Period	Specific description
1. Survey of present situation	To fully understand efforts concerning the accumulation and sharing of regional information in Otsuki City and Tsuru City	December 2011 to March 2012	- Survey of administrative measures and internet services - Interview targeting municipal employees and officials of residents' councils
2. System construction	To construct the system in detail to suit the region for operation	March to August 2012	- Define system requirements - Construct system - Developopeation system
3. Operation test	To conduct an operation test of the system	July 2012	- Creation and distribution of pamphlets and instructions for use - System operation test
4.Operation test evaluation	To reconstruct the system based on the results of interview of those who participated in the operation test	August to September 2012	- Evaluation through interview - System reconstruction - Revision of pamphlets and instructions for use
5. Operation	Conduct actual operation of the system	November to December 2012	- Appeal for use of the system - Distribution of pamphlets and instructions for use - System operation management
6. Operation evaluation	Evaluate the system based on the results of surveys by questionnaire and access analysis	December 2012	- Evaluation using web questionnaires and access analysis - Identification of improvement measures

TABLE II. COMPARISON OF CHARACTERISTICS OF EXISTING WEB SERVICES IN THE REGION FOR OPERATION WITH CHARACTERISTICS OF THE SYSTEM OF THIS STUDY

	Use of digital map	Aim	Means for accumulating regional information
Municipal official websites	Yes	Transmission of information by website operators	None
Tourist association websites	Yes	Transmission of information by website operators	None
e-machitown	No	Transmission of information by users, and accumulation and sharing of the information	Contributions to website by users
This system	Yes	Transmission of information by users, sharing and accumulation of the information, and exchange of the information	Contributions by users from computers or from portable information terminals using Twitter

information can be promptly deleted, and any user can also make an amendment using either of the two types of user screen, by making a comment in response to erroneous contributed information. Thus, this system has been developed keeping in mind the goals of reducing administrator burden as much as possible and developing a system which regular local residents in regional communities can also operate and manage.

V. OPERATION TEST AND OPERATION

According to the operation process in TABLEI, after the operation test and evaluation of the operation test of the social media GIS designed and developed in this study were conducted, actual operation was conducted.

A. Selection of the region for operation and anticipated users

The eastern part of Yamanashi Prefecture (Otsuki City andTsuru City) was selected as the region for operation of the system.This region adjoins the Tama region of the Tokyo Metropolis, and is dotted with tourist spots; therefore, it has many visitors from neighboring prefectures such as the Tokyo Metropolis.However, it has few web services which accumulate regional information.In order to examine the role and usefulness of the system in the region for operation, TABLEII compares the characteristics of three types of web service in the region for operation with characteristics of the system of this study.Even in cases where municipal and tourist association official websites use digital maps, the former are limited to just introducing facilities related to everyday life, various public facilities, and so forth, and the latter are limited to just presenting location information concerning tourist spots.The aim of both types of website is one-sided transmission of information by the operators; therefore, their aim differs from that of the system.Further, part of the aim of the website e-machitown is similar to the aim of the system; however, digital maps are not used in e-machitown.Based on the above, in terms of the fact that the system is oriented to transmission of information by users, and further, enables exchange of information between regions based on the accumulation and sharing of information, the roles of the system and existing web services in the region for operation can mutually complement each other.Further, the system demonstrates even more usefulness, in that since it is based on a digital map, it can provide a means for people inside and outside the region for operation to exchange information that relates more closely to the region.

Two types of user were anticipated as users in this study - users inside the region for operation, and users outside the region for operation.Users inside the region for operation use the system as a tool for accumulating and sharing information inside the region, and exchanging information with people outside the region for operation.Further, users outside the region for operation are mainly residents of the Tokyo Metropolis, and use the system as a tool for gathering information concerning the region for operation, and exchanging information with people inside the region for operation.

B. Operation test and operation test evaluation

As participants in the operation test, six people aged 18 years or older residing in the eastern part of Yamanashi Prefecture and two people aged 18 years or older residing in the Tama region of the Tokyo Metropolis were selected.The operation test was conducted for one week.It was observed that exchanges of information between regions were conducted during the operation test.In the interview held after the operation test, the following two areas for improvement were identified, so redevelopment of the system was carried out in regards to these two areas only.

1) Contributed information list pages and ranking pages were added so that the contributed information could be viewed on the pages other than the digital map page.

2) In the case where contributed information centers on information closely related to the everyday lives of local residents, and there are no contributions of information that is necessary to visitors from outside the region for operation, such as information concerning tourist information centers and transportation facilities, it is possible that the usefulness of the system will decrease.Therefore, with reference to official municipal and tourist association websites, regional information for visitors from outside the region for operation was inserted onto the digital map in advance.

Further, concerning contributions made from portable information terminals using Twitter, when tweets are made, location information obtained from a GPS using the portable information terminal is read and uploaded; therefore, the users cannot check location information themselves, and depending on the model of portable information terminal used, errors sometimes occur in the location information.Therefore, in this study, the use of twicca was recommended in both the operation test and the actual operation.Concerning twicca, after location information has been received from a GPS, it can be freely edited on the digital map; therefore, it is possible for the user to amend errors themselves.Further, even in the case where the information cannot be contributed immediately, if the user remembers the related location, they can add location information and contribute information afterwards.

C. *Operation*

In the eastern part of Yamanashi Prefecture, the region for operation, the motor vehicle is the main means for moving from place to place.Therefore, in the actual operation, those aged eighteen years and over, who are able to obtain a driver's license, were encouraged to use the system.TABLEIII is a breakdown of users during the operation period, which was approximately two months long.There were 20 users inside and 25 users outside the region for operation, giving a ratio of 4:5.Further, approximately 93% of users were university students or postgraduate students in their 20s; due to the nature of this system, the majority of users were of a generation that is considered proficient in the use of computers and portable information terminals.Further, users outside the region for operation were residents, workers or students in the Tama region of the Tokyo Metropolis.

Using the system, regional information can be contributed and viewed using various patterns from both computers and portable information terminals regardless of place andtime, and users outside the region for operation can make exchanges of information between regions using multiple methods, concerning information contributed by users inside the region for operation.Thus, encouraging use of the system suited to the everyday lives and preferences of each user enabled the effects of the system to be demonstrated to the fullest.

VI. Operation evaluation and identification of measures for improvement

In accordance with the operation process in TABLE I, the system was evaluated using the results of the questionnaire to the users described in the outline in TABLE III. Further, using analysis of access which employedlog data, exchange of information between regions, which was theaim of the system, was evaluated. Moreover, based on the results of the two evaluations, measures for improvement of the system were identified. As mentioned in Section V-C, due to the nature of the system, approximately 93% of the users were in their 20s, and the evaluation results are not based on users from a variety of generations; in this respect, the evaluation results lack generality to some degree.

TABLE III. Outline of users and questionnaire respondents

	Male	Female	Total	Questionnaire respondents
Inside the region for operation (Number of people)	12	8	20	18
Outside the region for operation (Number of people)	21	4	25	15
Total (Number of people)	33	12	45	33
Percentage (%)	73.3	26.7	-	73.3

A. *Evaluation of use of the system*

1) *Evaluation of operability of the system*

In order to operate the system over the long term, operability is important; therefore, as shown in Fig. 4, the operability of the system was evaluated.Concerning operationof the system, the percentage of respondents who answered "Easy" or "Fairly easy" was very high, at approximately 97%, a result which demonstrates the high level of operability of the system.This result is due to the fact that the design of the system was standardized with the design used in common SNS, and the fact that the system was designed such that it was easy to visually browse information, because rather than arranging a lot of information on one screen, contributed information was displayed on the digital map.

Concerning the operation of contributing from a portable information terminal using Twitter, a function that existing regional SNS do not have, evaluation responses diverged somewhat, with approximately 67% of respondents answering "Easy" or "Fairly easy", and approximately 27% answering "Average".Concerning this, the results reflect the fact that although the majority of users of the system were in their 20s and therefore considered to be proficient in using portable information terminals, some use Twitter routinely, while some do not.However in the section of the questionnaire where respondents could write their opinions freely, multiple responses mentioned that it was convenient that when users from outside the region for operation visited the region for operation, even if they did not have a grasp of the region geographically, a portable information terminal could be used to obtain location information from a GPS and contribute information using Twitter.Therefore, it can be said that for users who use Twitter routinely, Twitter was an effective means of contributing information to the system.

2) *Evaluation of functions unique to the system*

The button functions and the ranking function for evaluating contributed information which were described in Section IV-A were included in the system as unique features.Fig. 5 shows the results of evaluation of these unique functions of the system.For both functions, the percentage of respondents who answered "Necessary" or "Relatively necessary" was very high, at approximately 90% or more. This is because by using the buttons to reflect their intuitive thoughts, such as "I want to go there", users were able to easily indicate their intentions in response to information viewed.As outlined above, evaluation of the unique functions of the system was very high: therefore, this also contributed to the high evaluation of the overall operability of the system mentioned above.

3) *Frequency of use of the system*

Fig. 6 compares the frequency of everyday internet use as a means of obtaining regional information to the frequency of use of the system as a means of obtaining regional information during the operation period.For the former, approximately 70%

of respondents answered “Hardly used it at all” or “Did not use it”; however, for the latter, approximately 30% of respondents answered “Several times a week” and approximately 49% answered “Several times a month”, so it is clear that the system was used more frequently than the internet as a means of obtaining regional information.The evaluation was based on a limited operation period of approximately two months, and the reasons why the system was used in the manner described above are that during the operation period users mutually acknowledged each other and became interested in each other’s contributed information, and that particularly for users inside the region for operation, information contributed to the system was more closely related to local residents’ everyday lives and easier to take an interest in, compared to information on official municipal and tourist association websites.

4) Effects of use of the system

A high percentage of users inside the region for operation, approximately 80%, responded “I think so” or “I tend to think so” when asked whether viewing contributed information caused them to want to visit the place related to that information.Further, there were many cases where users employed the comment function to report actually visiting places which were the subject of contributed information, or to ask questions regarding a place indicated by contributed information.Meanwhile, among users outside the region for operation, the percentage who answered “I think so” or “I tend to think so” when asked whether viewing contributed information had provoked their interest in the region for operation was very high, at approximately 94%.There were cases where users outside the region for operation used the comment function to ask users inside the region for operation questions regarding details of information contributed.

B. *Evaluation of exchange of information between regions*

1) Outline of access analysis

In this study, log data collected during the operation period was used to perform access analysis.Thereby, inspection of whether or not information was exchanged between regions was performed, and as well, trends regarding the content of contributed information and users who made many contributions were clarified.This led to improvement of usability of the system and an increase in the access count.A Google Analytics API was included in the program developed in this study, and used to conduct access analysis.Google Analytics is free application software provided by Google, and is widely used as a tool for analyzing log data of websites.Google Analytics can be used just by writing the API in the program of the homepage of a website, and enables acquisition of users’ access log.

2) Access analysis results

TABLE IV shows the access analysis results, and TABLE V shows the results for classification of contributed information. From TABLE IV it can be seen that the access count from computers to the system of this study accounts for about 88% of the total access count, and the average number of page views and the average time spent viewing was as much as approximately twice the averages for access from portable information terminals, so it is clear that users mainly used the system via computers. However, in contrast to the fact that the access count from portable information terminals to the system was only about 12% of the total, the contribution count fromportable information terminals using Twitter was about 41%, only about 18% less than the contribution count fromcomputers. This difference is less than that of the difference between access counts for portable information terminals and computers. Based on this, it can be surmised that users felt little stress when contributing information from portableinformation terminals using Twitter, and therefore, although the access count from portable information terminals was low, the contribution count was high.Therefore, it can be said that the ease of contributing information using Twitter and the usefulness of having mashed up Twitter with the system were demonstrated.

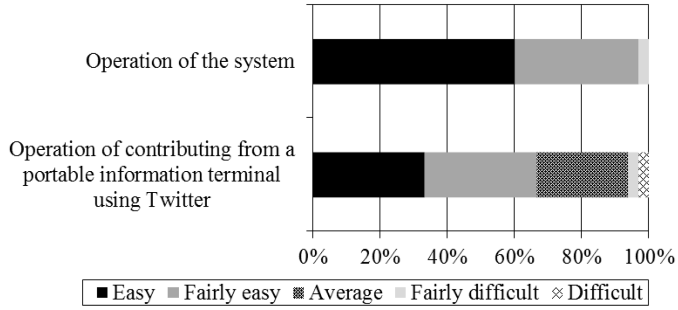


Fig. 4. Evaluation of the operability of the system (%)

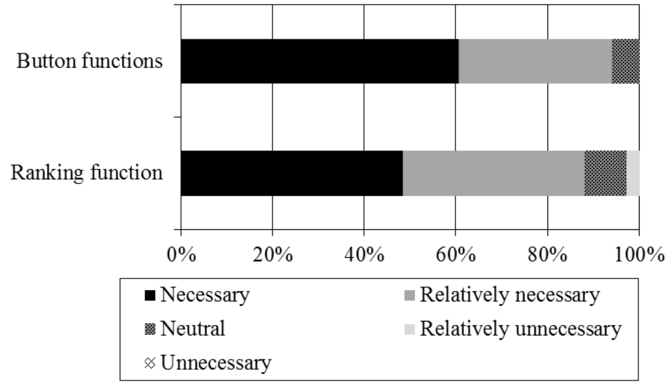


Fig. 5. Evaluation of unique functions of the system (%)

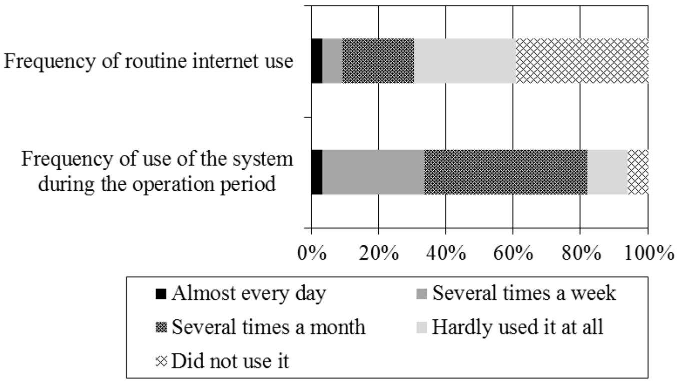


Fig. 6. Comparison of the frequency of using the internet to obtain regional information with the frequency of using the system to obtain regional information during the operation period (%)

TABLE IV. ACCESS ANALYSIS RESULTS

	Access count	Average number of page views	Average time spent viewing
Access from computers	338 (88.3%)	6.32	7 minutes, 34 seconds
Access from portable information terminals	45 (11.7%)	3.29	3 minutes, 36 seconds
Total access count	383	7.97	5 minutes, 5 seconds

TABLE V. RESULTS OF CLASSIFICATION OF CONTRIBUTED INFORMATION

Classification of contributed information	Places to eat and drink	Scenery	Other	Total
Number of contributions from computers (%)	15 (17.6)	25 (29.4)	10 (11.8)	50 (58.8)
Number of contributions from portable information terminals using Twitter (%)	20 (23.5)	7 (8.3)	8 (9.4)	35 (41.2)
Total number of contributions (%)	35 (41.1)	32 (37.7)	18 (21.2)	85 (100.0)

Moreover, as TABLE V shows, of the contributed information, about 41% concerned places to eat and drink, about 38% concerned scenery, and the remaining approximately 11% concerned various public facilities such as city halls and hospitals, and was information that is also published on official municipal websites. Investigating the content of each item of information contributed to the system in further detail, the information concerning places to eat and drink and scenery in particular noted small changes in central parts and in business patterns of places to eat and drink, and beautiful scenery in mountainous regions distant from urban districtsrespectively. Thus, it is clear that users within the region for operation contributed information that only local residents knew.

Further, the results of the evaluation of information contributed using button functions shows that the total number of points concerning all contributed information was 182 points, which is an average of 4.1 clicks per user.The breakdown is 150 points contributed using computers and 32 points contributed using portable information terminals, so there was a clear difference in the number of times button functions were used from computers and from portable information terminals.A cause of this is that photos were attached to the majority of information contributed from computers; therefore, it was easier to view the photos when using a computer, and the state of things related to contributed information, such as places to eat and drink and scenery, was conveyed to users in a more visual way.Further, it became clear that in the case of information contributed from portable information terminals using Twitter, a photo could not be viewed unless the user accessed the link, and therefore the button functions were not used as much as they were in the case of information contributed from computers.

3) Evaluation of exchange of information between regions, based on access analysis results

The case where the comment function was used as a means of exchanging information between regions occurred twelve times. An example of this kind of information exchange is exchanges of comments between users within and outside the region for operation in which users outside the region for operation who viewed photos attached to contributed information posted comments saying that they had actually visited the places after viewing them in the photos, and added new related information.Such exchanges took place several times.Similarly, the breakdown for cases where the button functions were used is that the “I didn’t know” button function was used 145 times (116 times from computers and 29 times from portable information terminals), and the “I want to go there” button function was used 37 times (34 times from computers and 3 times from portable information terminals).This shows that users within the region for operation used the button functions more, and use for both types of button functions was significantly greater from computers.However, it was confirmed that exchanges of information in which users outside the region for operation took interest in information contributed by users within the region for operation and used the“I want to go there” button function took place.Based on the above, it can be said that the system of this study functioned in accordance with its aims; as indicated in the previous section, the majority of information contributed was information known only to local residents of the region for operation; and the system fulfilled its purposes of changing the experience-based knowledge of local residents from implicit knowledge to explicit knowledge, accumulating and sharing the knowledge between people within the region,and exchanging the information with people from other regions.

C. Identification of measures for improvement

Based on the results of the questionnaires given to the users and the access analysis, measures for improvement of this system can be summarized into the following two areas:

1) Information classification and notification

Include a form which allows selection of categories showing interests and preferencesof users at the time of initial registration, and have users select categories.Similarly, when users contribute information, having them select categories in the manner described above would enable viewing of contributed information divided by categories.Further, the convenience of the system can be improved by including a contributed information notification function which notifies a user when information in a category of interest to the user has been contributed, or when there has been a comment in response to information contributed by the user.

2) Optimization of interface for smart phones

Since there were many contributions to the system from portable information terminals, the usefulness of the system can be improved by optimizing the interface for smart phones, which are becoming popular across a wide range of ages, so that the system can easily be used from smart phones when users are outdoors and cannot use a computer - for example, when they have gone out of the home or office for a while.

VII. Conclusions and Future Scopes

The conclusions of this study can be summarized into the following three areas:

1) Three web applications, that is, aWeb-GIS, an SNS, and Twitter, were integrated, and a social media GIS which enabled exchange of information between regions was designed and developed.The usefulness of the social media GIS was demonstrated in the three areas of interactivity of information transmission,easing of spatial and temporal limitations, and easing of limitations concerning continuing operation.Further, the eastern part of Yamanashi Prefecture was selected as the region for operation of the system, a survey of the existing situation was conducted, and then the system was developed in detail.Further, prior to actual operation being conducted, an operation test and an evaluation of the operation test were conducted, areas for improvement were identified, and the system was redeveloped.

2) Actual operation was conducted over approximately two months, with users aged eighteen or over from both inside and outside the region for operation, and due to the nature of the system, approximately 93% of the forty-five users were in their 20s.During the operation period information was contributed and viewed from both computers and portable information terminals, and exchanges of information between regions, in which the comment function and button functions were used in relation to contributed information, were conducted.

3) The system was evaluated based on the results of questionnaires given to the users and access analysis of log data.The questionnaire results showed the high level of operability of the system, the high frequency of use of the system during the operation period, and the large extent of the effects of use of the system, and it was clear that the high evaluation of the unique functions of the system also contributed to the high evaluation of the operability of the system.The access analysis results showed that while the access count from portable information terminals was no more than about 12%, the contribution count from portable information terminals was about 41%, accounting for slightly less than half of the total contribution count.Further, the results revealed that the majority of contributed information was information known only to local residents of the region for operation, and the system performed a role which was in accordance with its aims.

Future works are to add functions which support the measures for improvement identified in the previous section to the system, to design and develop a system such that people from various age groups can use it, and to operate and evaluate that system.A further future research task is to increase the usage track record of the system by operating it in other regions as well, and further increase the significance of usage of the social media GIS developed in this study.

Acknowledgment

In the operation of the social media GIS and the questionnaires of this study, enormous cooperation was received from those in the eastern region of Yamanashi Prefecture and the Tama region of the Tokyo Metropolis. We would like to take this opportunity to gratefully acknowledge them.

References

[1] Science Council of Japan:Regional Research Committee,“Toward to accumulate and applicate regional knowledge”, Tokyo, 2008.

[2] H. Itou, andK. Fukuyama,“A buisness-academia collaboration for regional developments on the basis of GIS”, Papers and Proceedings of the Geographic Information Systems Association of Japan,Vol.14, pp.551-554, 2005.

[3] S. Tsuboi, T. Sakai and S. Goto, “Research on the local rivitalization by application of GIS and its related factor -A case study of Kumagaya-uchiwa festival-.”,Gobal Environmental Studies,Vol.10, pp.41-47, 2008.

[4] J. Nuojua and K. Kuutti, “Communication based web mapping: A new approach for acquisition of local knowledge for urban planning”, Proceedings of the 12th International Conference on Entertainment and Media in the Ubiquitous Era, pp.136-140, 2008.

[5] Y. Shimizu, A. Kodama, A. Watanabe and R. Miyake, “A studyon sharing of reginal information as tacit knowledge in high-density commercial area: A trial of reginal informatization by demonstration mobile experiment in Akihabara”, Journal of Architecture and Planning, Vol.73, No.632, pp.2275-2280, 2008.

[6] B. Yu and G. Cai, “Facilitating participatory decision-making in local communities through map-based online discussion”, Proceedings of the Fourth International Conference on Communities and Technologies, pp.215-224, 2009.

[7] T. Umeta and M. Tomisawa, “Regional-oriented social network service”, Information Processing Society of Japan, Information Systems and the Social Environment -Research Reports,Vol.27, pp.69-76, 2006.

[8] Y. Hara, Y. Inaba, Y. Yoshiyuk, Natsuki, S. Motoseko and W. Yan, “Support for machizukuri on local collaboration and maintenance of local information by residential workshop”, Papers and Proceedings of the Geographic Information Systems Association of Japan,Vol.17, pp.497-500, 2008.

[9] K. Kitahara, H. Kanai, S. Urushibara and S. Kunifuji, “Communication-support system based on shared cooking ingredients for people living nearby”, Information Processing Society of Japan -Research Reports, GN:Groupware and Network Services, Vol.33, pp.13-18, 2009.

[10] C. Y. Chow, J. Bao and M. F. Mokbel, “Towards location-based social networking services”, Proceedings of the 2nd ACM SIGSPATIAL International Workshop on Location Based Social Networks, pp.31-38, 2010.

[11] T. Kirimura, K. Matsuoka andK. Yano, “Monitoring system of modern arcitectures and industrial heritages in Kyoto city usingWeb-GIS”, Papers and Proceedings of the Geographic Information Systems Association of Japan, Vol.17, pp.193-198, 2008.

[12] K. Soga, H. Fukada, H. Ichikawa and A. Abe, “Proposal of regional SNS cooperation map considering user behavior support”, Papers and Proceedings of the 70th Annual Conference of the Information Processing Society of Japan,pp.393-394, 2008.

[13] J. Nuojua, “Web map media: A map-based web application for facilitating participation in spatial planning”, Multimedia Systems, Vol.16, No.1, pp.3-21, 2010.

[14] N. Hosoya and K. Yamamoto, “Web-GIS based outdoor education program for elementaryschools”, Journal of Scio-Informatics, Vol.4, No.1, pp.49-62, 2011a.

[15] N. Hosoya and K. Yamamoto, “Web-GIS based outdoor education program for junior high schools”, World Academy of Science, Engineering and Technology, Vol.60, pp.316-322, 2011b.

[16] S. Kubota, K. Soga, Y. Sasaki, T. Miura, H.Takisawa, S.Takashi and A. Abe, “Development and operational evaluation of regional social networking service as public participation GIS”, Theory and Applications of GIS,Vol.20, No.2, pp.125-136, 2012.

[17] T. Yanagisawa and K. Yamamoto, “A study on information sharing GIS to accumulate local knowledge in local communities”,Theory and Applications of GIS, Vol.20, No.2, pp.61-70, 2012.

[18] H.Nakahara, T. Yanagisawa andK. Yamamoto, “Study on aWeb-GIS to support the communication of regional knowledge in regional communities : Focusing on regional residents' experiential knowledge”,Socio-Informatices, Vol.1, No.2, pp.77-92,2012.

[19] T. Fujisaka, R. Lee, and K. Sumiya, “Estimating influence regions of social events by geo-tagged micro-blogs analysis”, Institute of Electronics, Information and Communication Engineers, Proceedings of theSecond Forum on Data Engineering and Information Management, D7-4, 2010.

[20] C. H. Lee, “Mining spatio-temporal information on microblogging streams using a density-based online clustering method”, Expert Systems with Applications, Vol.39, No.10, pp.9623-9641, 2012.

[21] K. Fueda, andY. Kameda, “Search of the sightseeing information using Twitter”, Papers of the Conference of the Japanese Society of Computational Statistics, No.26, pp. 67-70, 2012.

[22] K. Sagawa, A. Hattori and H. Hayami, “An on-line map system for life Information at the time of the disaster”, Information Processing Society of Japan -Research Reports, GN:Groupware and Network Services, Vol.83, No.8, pp.1-7, 2012.

[23] R. Lee and K. Sumiya, “Measuring geographical regularities of crowd behaviors for Twitter-based geo-social event detection”, Proceedings of the 2nd ACM SIGSPATIAL International Workshop on Location Based Social Networks, pp.1-0, 2010.

[24] R. Lee, S. Wakayama and K. Sumiya, “Discovery of unusual regional social activities using geo-tagged microblogs”, World Wide Web, Vo.14, No.4, pp.321-349, 2011.

[25] Y. Hashimoto and M. Oka, “Statistics of geo-tagged tweets in urban area”, Journal of the Japanese Society for Artificial Intelligence, Vol. 27, No.4, pp.424-431, 2012.

[26] Z. Cheng, J. Caverlee and K. Lee, “You are where you tweet: A content-based approach to geo-locating twitter users”, Proceedings of the 19th ACM International Conference on Information and Knowledge Management, pp.759-768, 2010.

[27] S. Hiruta, T. Yonezawa and H. Tokuda, “Detection and visualization of place-triggered geotagged tweets”, Journal of the Information Processing Society of Japan, Vol.54, No.2, pp.710-720, 2013.

[28] Impress R&D, “Smartphone / mobile phone use pulse-taking 2013”, 2012, http://www.impressrd.jp/news/121120/kwp2013. (accessed February 16, 2013). (Website)

On The Performance of the Gravitational Search Algorithm

Taisir Eldos
Department of Computer engineering
College of computer engineering and Sciences
Salman bin Abdulaziz University
Al Kharj, Saudi arabia

Rose Al Qasim
Department of computer Engineering
Faculty of Engineering and Technology
Al Balqa Applied University
Amman, Jordan

Abstract—**Gravitational Search Algorithms (GSA) are heuristic optimization evolutionary algorithms based on Newton's law of universal gravitation and mass interactions. GSAs are among the most recently introduced techniques that are not yet heavily explored. An early work of the authors has successfully adapted this technique to the cell placement problem, and shown its efficiency in producing high quality solutions in reasonable time. We extend this work by fine tuning the algorithm parameters and transition functions towards better balance between exploration and exploitation. To assess its performance and robustness, we compare it with that of Genetic Algorithms (GA), using the standard cell placement problem as benchmark to evaluate the solution quality, and a set of artificial instances to evaluate the capability and possibility of finding an optimal solution. Experimental results show that the proposed approach is competitive in terms of success rate or likelihood of optimality and solution quality. And despite that it is computationally more expensive due to its hefty mathematical evaluations, it is more fruitful on the long run.**

Keywords*—*Optimization; Gravitational Search; Genetic Algorithms; Cell Placement

I. INTRODUCTION

GSA is a heuristic stochastic swarm-based search algorithm in the field of numerical optimization, based on the gravitational law and laws of motion. Like many other nature inspired algorithms, it needs refinements to maximize its performance in solving various types of problems. In addition to the problem encoding that sometimes can be a challenge, fine tuning its parameters play a significant role balancing the search time versus solution quality. This algorithm is relatively recent and not heavily explored.

Cell placement is one of four consecutive steps in physical design process of VLSI circuits, namely: partitioning, placement, routing and compaction. In the placement stage, the description of the physical layout of the chip is introduced, by assigning geometric coordinates to the cells. The objective of the placement algorithm is to find a layout that minimizes a cost function, whose major part is the area, but quite often involves the aspect ratio, to make the chip as close to square as possible and hence increase the die yield.

II. LITERATURE REVIEW

Approaches to solve cell placement problem are generally classified into two classes; constructive and iterative improvement methods. Several heuristic optimization strategies for solving placement problem have been implemented via a set of diversified algorithms; evolution-based placement like Genetic Algorithms [5] and Simulated Annealing [6], and a comprehensive summary of those strategies is presented in [1].

Gravitational Search Algorithms (GSAs) are novel heuristic optimization algorithms introduced in [2], and researched in the past few years, as a flexible and well-balanced strategy to improve exploration and exploitation methods. In [3], the binary gravitational search algorithm was developed to solve different nonlinear problem. A new multi-objective gravitational search algorithm was proposed in [4]. The GSA shows satisfactory results for solving many problems in a various applications; Solving Symmetric Traveling Salesman Problem [7], solving the flow shop scheduling problem [8], in feature selection [9], image enhancement [10], solving DNA sequence design problem [11], and optimize the filter modeling parameters [12]. A hybrid algorithm was derived from both Genetic Algorithms and Gravitational Search Algorithm for feature set selection [13].

In this paper, we enhance our implementation of the gravitational search technique, to solve the cell placement, with the intention compare its performance with well know evolutionary algorithms in future work. The results show that the algorithm can improve the solution quality in a reasonable amount of time. This paper is organized as follows: Section 2 gives a formal description of the GSA theory, Section 3 gives a brief description of the cell placement problem, section 4 demonstrates the proposed gravitational search algorithm for cell placement, In section 5 we discuss the performance of this algorithms in solving standard problems as compared to the well know genetic algorithm, and section 6 wraps up our work.

III. METHODOLOGY

The Gravitational Search Algorithm (GSA) was proposed by Rashedi [2], as a simulation of Newton’s gravitational force behaviors. In this algorithm, possible solutions of the problem in hand are considered as objects whose performance (quality) is determined by their masses, all these objects attract each other by the gravity force that causes a global movement of the objects towards the objects with heavier masses. The position of each object corresponds to a solution of the problem, and inertial masses are determined by a fitness function. The heavy masses, which represented a good solutions, move more slowly than lighter ones, this represents the exploitation of the algorithm.

The GSA starts with a set of agents, selected at random or based on some criteria, with certain positions and masses representing possible solutions to a problem, and iterates by changing the positions based on some values like fitness function, velocity and acceleration that gets updated in every iteration. To relate those values and parameters, let us demonstrate the relations among them.

In a system with N agents, the position of the ith agent is defined as:

$$X_i = \left(x_i^1, \dots, x_i^d, \dots, x_i^n\right) \; for \quad i = 1,2,\dots,N \tag{1}$$

Where x_i^d present the position of the ith agent in the dth dimension, and ***n*** is dimension of the search space.

At the time t a force acts on mass i from mass j. This force is defined as follows:

$$F_{ij}^d = G(t)\frac{M_{pi}\,(t)\times M_{aj}(t)}{R_{ij}+\varepsilon}(x_j^d(t) - x_i^d(t)) \tag{2}$$

Where M_{aj} is the active gravitational mass of agent j, M_{pi} is the passive gravitational mass of agent i, G(t) is gravitational constant at time t, ε is a small constant, and $R_{ij}(t)$ is the Euclidian distance between two agents i and j:

$$R_{ij}(t) = \left\|X_i(t).X_j(t)\right\| \tag{3}$$

The total force acting on $mass_i$ in the dth dimension in time t is given as follows:

$$F_i^d(t) = \sum_{j\epsilon Kbest, j\neq i}^{N} rand_j F_{ij}^d\,(t) \tag{4}$$

Where $rand_j$ is a random number in the interval [0, 1], K best is the set of first K agents with the best fitness value.

The acceleration related to mass i in time t in the dth dimension is given as follows:

$$a_i^d = \frac{\mathrm{F_i^d(t)}}{\mathrm{M_{ii}(t)}} \tag{5}$$

Where M_{ii} is the inertial mass of ith agent.

The next velocity of an agent could be calculated as a fraction of its current velocity added to its acceleration. Position and velocity of agent is calculated as follows:

$$v_i^d(t+1) = rand_i \; v_i^d(t) + a_i^d(t) \tag{6}$$

$$x_i^d(t+1) = x_i^d(t) + v_i^d(t+1) \tag{7}$$

Where $rand_i$ is a uniform random variable in the interval [0, 1].

Gravitational constant, G, is initialized at the beginning of the search and will be reduced with time to control the search accuracy as follows:

$$G(t) = G_0 e^{-\alpha\frac{t}{T}} \tag{8}$$

Where T is the number of iteration, G_0 and α are given constant.

The gravitational mass and the inertial mass are updated by the following equations:

$$M_{ai} = M_{pi} = M_{ii} = M_i, \qquad i = 1, 2, \dots., N \tag{9}$$

$$m_i\,(t) = \frac{fit_i\,(t) - worst(t)}{best(t) - worst(t)} \tag{10}$$

$$M_i\,(t) = \frac{m_i(t)}{\sum_{j=1}^{N} m_j(t)} \tag{11}$$

Where $fit_i(t)$ represent the fitness value of the agent i at time t, and, worst(t) and best(t) are given as follows for a minimization problem:

$$best(t) = \min_{j\in\{1,..,N\}} fit_j(t) \tag{12}$$

$$worst(t) = \max_{j\in\{1,\dots,N\}} fit_j(t) \tag{13}$$

IV. Cell Placement Problem

We use the Normalized Polish Notation (RPN) [14] to describe any arrangement representing a possible solution; for n cells, a string with n modules (cells) and n-1 operators of the * or + type, to mean above or next to. As an example, the string (2 3 * 1 + 4 5 + 6 7* + *) is an encoding for the arrangement in Figure 1. Here, relaxed means the case where the area is that of the minimal rectangle enclosing the cells, while the restricted means the case where the area is that of the minimal square enclosing the cells.

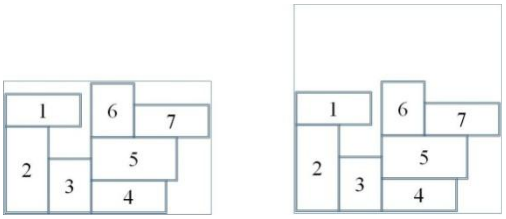


Fig. 1. (a) Relaxed and (b) Restricted area

Such a configuration is an agent in gravitational search algorithm; new agents are generated from the existing ones by applying certain operators which are described in [14] and [15]. New solutions are assigned fitness values that reflect their quality. We propose the following fitness measure:

$$\mathrm{F} = \alpha\frac{\mathrm{A}}{\mathrm{SL}} + (1-\alpha)\frac{\mathrm{S}}{\mathrm{L}} \tag{14}$$

Where L and S are the long and short sides of the rectangle enclosing all the cells and A is the algebraic sum of the areas of all cells regardless of the placement, and the product SL represents the area associated with the solution. The factor α is a number between 0 and 1, introduced to dictate the relative significance of the aspect ratio to the actual area; to favor square arrangements we use smaller values of α. If α=1 then aspect ratio is not optimized.

V. Gravitational Search Algorithm Adaptation

Cell placement can be viewed as a two-dimensional bin packing problem, where the goal is to arrange a number of cells with different sizes in a way that reduces the area enclosing them and producing near square die while providing enough space for efficient routing. In this sense, we propose a new algorithm for cell placement problem by means of GSA, in which each mass will be an agent looking for an optimal solution in the search space.

Since cell placement needs meet simultaneously several constraints, it is difficult to be solved by the traditional GSA. For this reason, the definition of distance between solutions (positions) and their update are modified as will be shown in the following procedure:

$$mass_i(t) = \frac{fitness_i(t) - worst}{\sum_{j=1}^{N}(fitness_j(t) - worst)} \tag{15}$$

$$force_{ij}(t) = G(t)\frac{mass_i(t) \times mass_j(t)}{distance_{ij}(t) + \epsilon}, \qquad \text{where €=0.1} \tag{16}$$

$$G(t) = G_{ini}\left(1 - \frac{iteration}{total\ iteration}\right), \qquad \text{where } G_{ini}=100 \tag{17}$$

$$totalforce_i(t) = \sum_{j=1}^{N} rand_j \times force_{ij}(t) \tag{18}$$

$$acceleration_i(t) = \frac{totalforce_i(t)}{mass_i(t)} \tag{19}$$

$$velocity_i(t+1) = rand_i \times velocity_i(t) + acceleration_i(t) \tag{20}$$

$$propapility_i = |\tanh\,(velocity_i(t+1)| \tag{21}$$

However, the position updating equation (7) cannot be applied in our case, because we are working in a string form to present the solution. Therefore, Rashedi [3] proposed the Binary GSA for nonlinear problem, which has the same formulation presented above, but with a different equation for updating the position of each agent. In order to update our solution, the formulation of binary GSA is used as shown in step 3.6. However, the stopping criteria can be based time budget or number of iterations, or reaching a target fitness or cost function (area in cell placement), or an improvement rate less than a threshold.

VI. The Algorithm Outline

The gravitational search algorithm is outlined as follows:

```
1.  Generate initial population of N agents at random
2.  Compute G(t), Best Fitness and Worst Fitness
3.  For each agent i, do:
    3.1.  Evaluate $Fitness_i$
    3.2.  Evaluate $Mass_i$
    3.3.  Evaluate Force of $Mass_i$
    3.4.  Evaluate Acceleration of $Mass_i$
    3.5.  Update Velocity of $Mass_i$
    3.6.  Find new Position of $Agent_i$
              If ($Probability_i$ > Threshold)
              {
                  If ($Rand_i$ < $Probability_i$)
                  Then Pair $Solution_i$ with the Best
                    Fit Solutions
                  Else Impose some minor change to
                    $Solution_i$
              }
4.  If Stopping Criteria Not Met, Go To 2 Else Stop
```

VII. Results

We carried two kind of tests; one on standard benchmark problems to evaluate the quality of the solutions, and another on artificial problems with known optimal solutions to measure the possibility of finding the optimal solution. The algorithm has achieved good results regarding the solution quality and success rate in finding optimal solution.

In the first study, three MCNC benchmarks; Xerox with 10 cells, Ami33 with 33 cells, and Ami49 with 49 cells, selected from MCNC and tested. Table1 1, 2, and 3 summarize the results of running the two algorithms on one of the benchmark problems in Table 1. For each case, 10 runs with different initial solutions are performed, for fixe number of iterations each. The number of iterations is set to a value proportional to the problem size. Clearly, GSA outperformed GA in the best, worst and mean waste as a measure all the time, and the aspect ratio most of the time.

TABLE I. Performance Comparison: (Xerox 10), 15000 Iterations

	GA	GSA
Best wasted area, Aspect	5.9 %, 1.83	4.2 %, 1.63
Worst wasted area, Aspect	8.3 %, 1.22	6.7 %, 1.05
Mean wasted area	7.0 %	5.4 %

TABLE II. performance comparison: (ami33), 30000 iterations

	GA	GSA
Best wasted area, Aspect	8.6 %, 2.12	6.1 %, 1.45
Worst wasted area, Aspect	14.2 %, 1.78	9.1 %, 2.11
Mean wasted area	11.2 %	7.2 %

TABLE III. Performance comparison: (ami49), 50000 iterations

	GA	GSA
Best wasted area, Aspect	13.1 %, 2.4	8.1 %, 1.56
Worst wasted area, Aspect	18.2 %, 2.3	13.2 %, 1.21
Mean wasted area	16.1 %	9.8 %

Fig. 2 shows the progress of the two algorithms over time; wasted area of the best solution in hand after multiple thousands of iterations. Same initial set of solutions (with 42% waste best case) evolved relatively at the same rate in the first few thousands of iterations, and then the GSA starts having better progress.

The second test is carried out on three artificial problems with 10, 20 and 30 cells of known optimal solutions, where a square is split into smaller squares and rectangles to generate instances with the target size, as shown in Figure3. Both GSA and GA are run 6 times with different initial solutions.

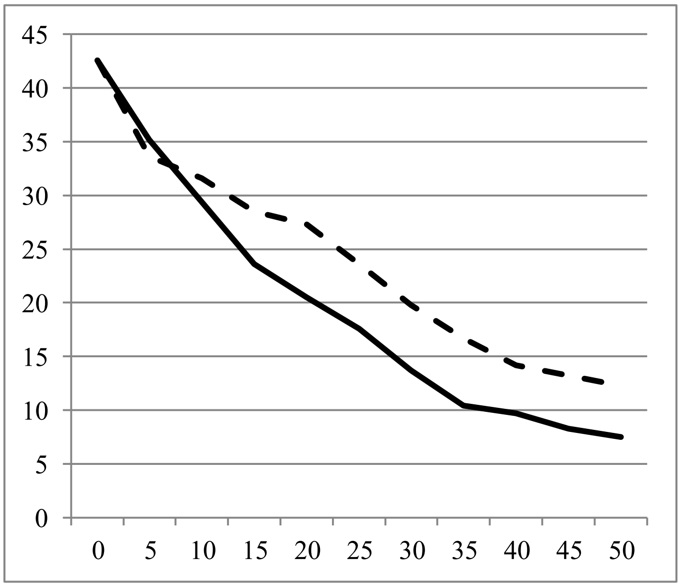


Fig. 2. Search Progress; Waste area for Ami49 versus Iterations (GSA solid, GA dashed)

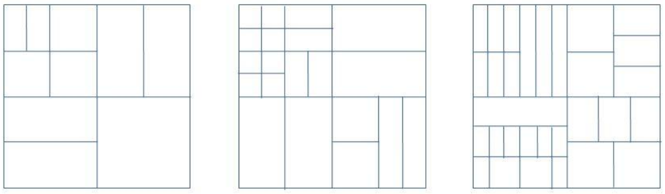


Fig. 3. Artificial Instances for Know Optimal Solutions

The algorithm is brought to stop if an optimal solution is achieved or the number of iterations equals 15000, 30000 and 50000 for the 10, 20 and 30 cells respectively. Table 4 shows the success rate or the likelihood of optimality. Again, GSA beats GA in for the small medium and large size, with significant outperformance of 100% in the 10 cells instance

TABLE IV. SUCCESS RATE OF ARTIFICIAL PROBLEMS (GSA VS. GA)

		Success Rate	
No. of Cells	No. of Iterations	GSA	GA
10	15,000	6 out of 6	4 out of 6
20	30,000	3 out of 6	1 out of 6
30	45,000	2 out of 6	1 out of 6

A major drawback of thi technique is its computational requirement; each iteration needs to many computations compared to other evolutionary algorithms like genetic algorithms for example. However, the effectiveness of this search and its balance between exploration and exploitation overcome this drawback. Table 5 shows the time taken by the GSA and GA to solve a 20-cell artificial instance with known optimal solution, running with same initial set of solutions on a personal computer with moderate specs. Both algorithms are made to stop when they reach a solution with some target quality; 5%, 10% 15% and 20% of wasted area relative to the optimal area.

TABLE V. TIME REQUIREMENTS FOR 20 CELLS INSTANCE

	Time (Minutes)	
Wasted Area	GSA	GA
5%	42.3	54.3
10%	34.8	42.9
15%	31.2	30.6
20%	23.1	23.8

VIII. CONCLUSION

The GSA power of solving a relatively complex problem, such as Cell Placement, is investigated using both benchmark and artificial instances with various sizes. Comparative tests have shown that GSA outperforms GA as a well known evolutionary algorithm, in terms of solution quality, i.e. the wasted area of the best configuration, aspect ratio, and the likelihood of finding optimal solutions. It is quite significant to note that although iterations take longer time in GSA compared to GA, the total time required to achieve a target solution quality is less when we target higher quality solutions. While the two algorithms take nearly the same amount of time to find decent solutions, targeting high quality solutions; 5% waste or less, can be achieved in 75% of the time with GSA. After the first few thousands of iterations, GSA outperforms GA by 10% to 40% in terms of wasted area.

REFERENCES

[1] K. Shahookar and P. Mazumder, "VLSI Cell Placement Techniques." ACM Computer Survey, vol.23, no. 2, pp. 143–220, June 1991.

[2] E.Rashedi, H.Nezamabadi-pour, and S.Saryazdi, "GSA: A Gravitational Search Algorithm." Journal of Information of Science 179, 2232-2243, 2009.

[3] E. Rashedi, H. Nezamabadi-pour, and S. Saryazdi, "Binary Gravitational Search Algorithm." Springer Science + Business Media B.V. 2009

[4] H. R. Hassanzadeh, M. Rouhani, "A Multi-Objective Gravitational Search Algorithm." International Conference on Computational Intelligence, Communication Systems and Networks. 24, pp117-122, 2010.

[5] T. W. Manikas, M.H. Mickle, "A Genetic Algorithm for Mixed Macro and Standard Cell Placement." The 45th Midwest Symposium on Circuits and Systems, vol. 2, pp 115 - 118, vol.2, August 2002

[6] G. Nan, M. Li, D. Lin, J. Kou. Adaptive Simulated Annealing for Standard Cell Placement. Springer, 2005 Advances in Natural Computation Vol. 3612, pp 943-947, 2005

[7] A. R. Hosseinabadi, M. yazdanpanah and A. S. Rostami, A New Search Algorithm for Solving Symmetric Traveling Salesman Problem Based on Gravity, World Applied Sciences Journal 16 (10): pp 1387-1392, ISSN 1818-4952, 2012

[8] Gu W X, Li X T, Zhu L, "A gravitational search algorithm for flow shop schduling. CAAI Transactions on Intelligent Systems". 5(5): 411-418, 2010.

[9] J.P. Papa, A. Pagnin, S.A. Schellini, A. Spadotto. Feature selection through gravitational search algorithm. Acoustics, Speech and Signal Processing (ICASSP), IEEE International Conference on, May 2011, pp: 2052 – 2055, 2011.

[10] W. Zhaoa, "Adaptive Image Enhancement based on Gravitational Search Algorithm." Procedia Engineering 15, pp. 3288 – 3292 Published by Elsevier Ltd. 2011.

[11] J. Xiao, Z. Cheng, "Theories and Applications DNA Sequences Optimization Based on Gravitational Search Algorithm for Reliable DNA computing." Sixth International Conference on Bio-Inspired Computing. IEEE Computer Society, 2011

[12] Rashedi E, Nezamabadi-pour H, Saryazdi S, "Filter modeling using gravitational search algorithm. Engineering." Applicaitons of Artifical Intelligence, 24, pp. 117-122, 2011

[13] M. Omar and J. Al-Neamy, "Hybrid Gravitational Search Algorithm and Genetic Algorithms for Automated Segmentation of Brain Tumors Using Feature_based Symmetric Ananlysis", (IJCSIS) International Journal of Computer Science and Information Security, Vol. 11, No. 5, May 2013.

[14] D. F. Wong, and C. L. Liu, A New Algorithm for Floorplan Design, Proc. DAC, pp.101–107,1986.

[15] J. P. Cohoon, S. U. Hedge, W. N. Martin, and D. S. Richards, “Distributed genetic algorithms for the floorplan design problem,” IEEE Trans. Computer-Aided Design, vol. 10, pp. 483–492, Apr. 1991.

Evaluation of Human Emotion from Eye Motions

Vidas Raudonis
Department of Control Technologies, KTU
Kaunas, Lithuania

Agne Paulauskaite - Taraseviciene
Department of Business Informatics, KTU
Kaunas, Lithuania

Gintaras Dervinis
Department of Control Technologies, KTU
Kaunas, Lithuania

Gintare Kersulyte - Raudone
Department of Control Technologies, KTU
Kaunas, Lithuania

Andrius Vilkauskas
Faculty of Mechanical Engineering and Mechatronics,
KTUK aunas, Lithuania

***Abstract*— The object of this paper is to develop an emotion recognition system that analysis the motion trajectory of the eye and gives the response on appraisal emotion. The emotion recognition solution is based on the data gathering using head mounted eye tracking device. The participants of experimental investigation were provided with a visual stimulus (PowerPoint slides) and the emotional feedback was determined by the combination of eye tracking device and emotion recognition software. The stimulus was divided in four groups by the emotion that should be triggered in the human, i.e., neutral, disgust, exhilaration and excited. Some initial experiments and the data on the recognition accuracy of the emotion from eye motion trajectory are provided along with the description of implemented algorithms.**

Keywords—Emotional status; eye tracking; human computer interaction; virtual human

I. INTRODUCTION

Recently emotional analysis has become an important line of research in human computer interaction (HCI), virtual human [1] or creating domestic robots [19] emotional agents [11]. Emotion recognition is one type of sentiment analysis that focuses on identifying the emotion in images of facial expression, measurements of heart rates, EEG, etc [3], [20]. Automatic recognition of emotion from the natural eye motions is an open research challenge due to the inherent ambiguity in eye motion related to certain emotion [2]. Various techniques have been proposed for emotion recognition. They include an emotion lexicon, facial expression models, EEG measurements with combination of knowledge-based approaches.

Humans interact with each other mostly by speech. However, regular human speech is often enriched with certain emotions. Emotions are expressed by visual, vocal and other physiological ways. There is evidence that certain emotion skills are part of what is called human intelligence that helps to understand better each other [18]. There exist an old saying “the eyes are the mirror of the human soul”. The facial expressions is only one of the ways to display an emotion, the emotion can be deduced from the eye gaze [4], [5]. If there exist a need for more affective human-computer interaction, recognizing of the emotional state from his or her face could prove to be an invaluable tool. The accurate emotion recognition is important in the field of the virtual human design, which should be with lively facial expression and behaviours, present motivated responses, to environment and intensifying their interaction with human users.

Eye-tracking can be especially useful investigating the behaviour of the individuals with physical and psychological disorders. Such studies typically focus on the processing of emotional stimuli, suggesting that eye-tracking techniques have the potential to offer insight into the downstream difficulties in everyday social interaction which such individual’s experience [6], [7]. Studies such as [5], [8], [10] describe how eye analysis can be used to understand human behaviour, the relationship between pupil responses and social attitudes and behaviours, and how it might be useful for diagnostic and therapeutic purposes.

The emotional part closely correlates to the eye based HCIs [10], [11], [12]. For example it is known, that increases in the size of the pupil of the eye have been found to accompany the viewing of emotionally toned or interesting visual stimuli. Emotional stimuli may also be used to attract the user’s attention and then divert to a control scheme of a HCI [9], [12]. E.g., eye tracking can be used to record visual fixation in nearly real-time to investigate whether individuals show a positivity effect in their visual attention to emotional information [13], [16], [17]. Different viewing patterns can be detected using strongly stimulant pictures, as in the study [14], [15].

This paper describes a system that uses a machine learning algorithm such as artificial neural network in order to detect four emotions: (a) neutral, (b) disgust, (c) funny and (d) interested from the eye motions. The text bellow presents our work on the development of emotional HCI. The proposed emotional status determination solution is presented based on the data gathering from visual sensors, providing the participants with a visual stimulus (strongly stimulant slides). Some initial experiments and the data on the recognition accuracy of the emotional state based on the gaze tracking are provided along with the description of implemented algorithms.

II. TECHNICAL IMPLEMENTATION OF EMOTION RECOGNITION METHOD

Our gaze tracking device is shown in the figure 1 (left). It consists of one mini video camera that is directed to the user's eye and it records the eye images. Next to the camera the infrared (IR) light source is attached which illuminates the users eye. The camera is design to capture the user's eye in the IR light. Such illumination does not disturb the user, because it is invisible for the naked human eye and gives mostly stable IR luminosity. The eye image captured in the near infrared light always shows dark (black) pupil that is caused by the absorption of infrared light of the eye tissues. Thus, the task of pupil detection can be limited or simplified to the searching goal of the dark and round region in the image. The camera and IR led is fixed on the ordinary eye glasses. Video camera is connected to the personal computer through USB port. Computer records the eye images, estimates the pupil size and coordinates, and draws the attention maps.

The constant adaptation of the eye pupil to the random variation of the illumination condition and different pupil sizes between user groups are the main challenges for any gaze tracking algorithm. In this work we propose the algorithm which is differ from well known method such as Circular Hough Transform or Viola-Jones. First method is based on the voting procedure that is carried out in a parameter space, from which circle candidate is obtained as local maxima in so called accumulator space. When the pupil size or circle diameter is unknown, the algorithm based on the Hough transform computes the accumulator space for every possible circle size and the circle candidate is obtained as global maxima. Such computation task takes relatively a lot of time. Viola-Jones based object tracking algorithm is capable to tracking the object of any shape and size based on the learned features. Unfortunately, it cannot be used when accurate measurements of the pupil size are required. Proposed pupil detection algorithm is based on adaptive thresholding of grayscale image. It enables the precise detection of pupil in the different layers of gray color, regardless of how the lightening is changing. The diameter measurement of the dark region is compared with the limits of the possible minimal and maximal pupil size.

Our gaze tracking algorithm finds the rough pupil center in the iterative manner and it executes the logical indexing on the gray level image using certain threshold of grayness value, which is variable (adaptive). The algorithm is preceded in two major steps. The rough pupil center is obtained in the first, and the accurate coordinates and the diameter of the pupil is obtained in the next step. The detection of the pupil is executed in the region of interest (sliding window) that is defined by three parameters Length, Width and the center coordinates (Cx, Cy) .All logical indexing operations are executed in the region of interest. At the detection beginning, the center of eye image is used as starting position for the region. All other positions are defined by located pupil center in the last frame. The values of all pixels which are higher than threshold are equalized to one, otherwise to zero. The threshold Θ is increased or reduced from the default gray level value Θ0 according to certain conditions which are defined by the current measured diameter r of the object of interest. The threshold Θ is increased by step ΔΘ, when current diameter of the object is smaller than the possible limits of the pupil size [Rmin Rmax] and otherwise Θ is decreased by step ΔΘ if these limits are exceed. Where Rmin – is the minimal pupil size and Rmax – is the maximal pupil size. These parameters are measured in image pixels. The variation limits of the pupil size of the human eye are taken according to analytical research [25]. The threshold value does not change if current measurement r is between limits. The rough pupil center (coordinates Cx and Cy) is computed in the next step. Used notations: N, M – the number of columns and rows of the eye image, d(i,j) – Euclidean distance between two candidate points, x, y – the coordinates of the candidate pixels, flag and count – is used for iteration purposes and σ - standard deviation of Euclidean distances. More about eye tracking algorithm is published in [12]. The schematic pseudo code of the proposed eye tracking method is shown in the tables 1 and 2.

TABLE I. THE PROPOSED ALGORITHMS FOR EYE PUPIL TRACKING AND REGISTRATION

Pupil detection based on adaptive gray level threshold

```
 //The pupil center is extracted in the grayscale image of the eye $\Gamma(u,v)$, where $u,v \in N,M$
while ($flag$ = 0) do
  //Collect candidate points
    $count = count + 1, k = 0;$
    for $u, v \in N, M$ do
      if $\Gamma(u, v) \le \Theta$ then
          $k = k + 1, x_k = u, y_k = v;$
    for $i, j \in k$ do
        $d(i,j) = \sqrt{(x_i - x_j)^2 + (y_i - y_j)^2}$
  //Make measurements of the point cloud
    $r = \max_{i \in k}(\max_{j \in k}(d(i,j)))$
    $C_x, C_y \leftarrow x(k), y(k)$
  //Verify the conditions
    if $R_{min} < r < R_{max}$ then, $flag = 1$,
      else if $R_{min} > r$ then, $\Theta = \Theta + \Delta\Theta$
        else if $R_{max} > r$ then, $\Theta = \Theta - \Delta\Theta$
```

TABLE II. THE PSEUDO CODE FOR ACCURATE DETECTION OF PUPIL CENTER

Accurate pupil center detection

```
//Accurate pupil center $\boldsymbol{C'} = (C'_x, C'_y)$of the point cloud $x_k$ and $y_k$ is computed using nonlinear least squares approach
//Perform data filtration on the candidate points
  //Collect candidate points
    $d_i = \sqrt{(\bar{x} - x_i)^2 + (\bar{y} - y_i)^2}$
    $\bar{d}, \sigma_d \leftarrow d(i), n = 0$
    for $i \in k$ do
      if $\bar{d} - 3\sigma_d \le d(i) \le \bar{d} + 3\sigma_d$ then
        $n = n + 1, x'_n = x_i, y'_n = y_i$
  //Fit smallest surrounding circle to the filtrated data
$\boldsymbol{u} = (C'_x, C'_y, r'), \mathbf{X} = \begin{pmatrix} x'_1 & \cdots & x'_n \\ y'_1 & \cdots & y'_n \end{pmatrix}, \mathbf{X} = R^2$
$f_j(C'_x, C'_y, r') = \left\|\mathbf{C}' - \mathbf{X}_j\right\|^2 - r^2$
$C'_x, C'_y, r' \leftarrow \sum_{j=1}^{m} f_j(\tilde{\boldsymbol{u}})^2 = \min_u \sum_{j=1}^{m} f_j(\boldsymbol{u})^2$
```

The computational actions of the proposed algorithm, taken for pupil detection, are shown in the figure 1. The threshold Θ is iteratively changed and the distance between two extremities of the selected pixels is measured. These pixels usually appeared in opposite direction from each other. The pixels are marked with lighter color in figure 1. The threshold value is decreased from 130 to 120 gray level and the same measurements are applied to the new selected pixels. Gray threshold is reduced until the distance between extremities is in the possible variation range of the pupil size. The grayness value was reduced to 90 of the gray level in this case. If selected pixels do not satisfy the predefined condition then Θ is changed. The threshold is increased when the distance between two extremities is smaller than minimum limit of the pupil size variation and decreased when this distance exceeds predefined maximum limit. The least square method is applied to selected pixels and the new accurate pupil center is obtained. The possibility of the proposed algorithm to change the threshold value adaptively overcomes several detection difficulties, such as, the variation of an ambient luminosity, constantly adapting pupil size and noise.

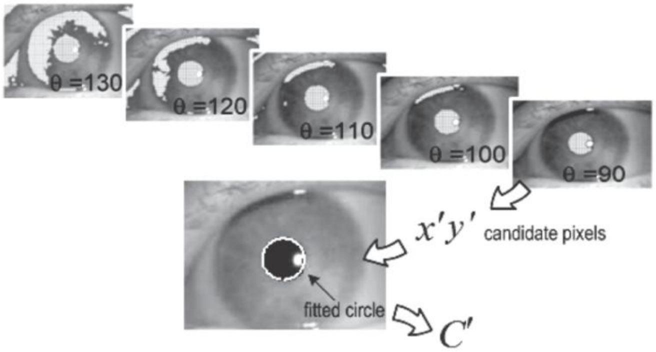


Fig. 1. The process of the adaptively changing threshold

Most eye movement is executed without awareness, i.e., there is no voluntary control. A sufficient amount of studies worldwide prove an interrelation between pupil size, pupil motions and a person's cognitive load or stress. The eye movements and the size of eye pupil strongly depend on the various factors of the environment and mental state of person. To recognize in which mental state is the person, the authors of this article have developed an eye pupil analysis system that is based on application of artificial neural network (ANN). The relation between measured inputs and the emotional human states is not known precisely, i.e., is it linear or nonlinear. Therefore, in the problems when linear decision hyper-planes are no longer feasible, an input space is mapped into a feature using hidden layers of the neural network. The mathematical model based on ANN is selected in order to construct a non-linear classifier.

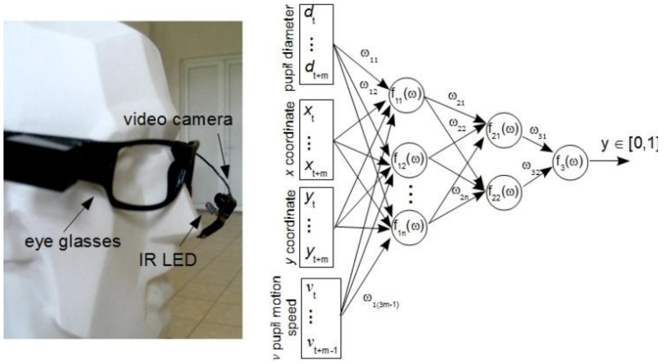


Fig. 2. The ANN model and hardware implementation of experimental investigation

Our experimental emotion detection system is illustrated in figure 2. In addition to the gaze tracking hardware/software the system runs the proprietary real-time emotion analysis toolkit based on an Artificial Neural Networks. We have implemented a 3 layer ANN: consists of 8 neurons, the second of 3 neurons and the output layer of 1 neuron. ANN networks have a variable input number and are trained based on 3 features: the size of the pupil, ant the position of the pupil (coordinates x, y) and motion speed ν of the eye. For each emotional state we develop different neural network. The artificial neural network can be described using the following formulas:

$$X = ([d_t..d_{t+m}][x_t..x_{t+m}][y_t..y_{t+m}][v_t..v_{t+m}]); \quad (1)$$

$$\mathrm{y}'_\mathrm{j} = \mathrm{f}\left(\textstyle\sum_{\mathrm{i}=1}^{3\mathrm{m}} \mathrm{X}_\mathrm{i}\omega_{\mathrm{ji}}\right); \quad (2)$$

$$\mathrm{y}''_\mathrm{k} = \mathrm{f}\left(\textstyle\sum_{\mathrm{j}=1}^{\mathrm{n}} \mathrm{y}'_\mathrm{j}\omega_{\mathrm{kj}}\right); \quad (3)$$

$$\mathrm{y} = \mathrm{f}(\textstyle\sum_{\mathrm{k}=1}^{2} \mathrm{y}''_\mathrm{k}\omega_2); \quad (4)$$

where, t – is the current sample, X – the input vector of the artificial neural network, y – output, ω – weights of the neural network, d – is the diameter of the recognized pupil, $\mathrm{x_t}$ and $\mathrm{y_t}$ are the coordinates of the pupil center and $\mathrm{v_t}$ is the speed of eye pupil motion. Decision is made by selecting maximal value from four outputs of the artificial neural networks.

III. Experimental Setup, Investigation and Results

At this initial stage of the evaluation we have chosen to analyze 4 very common emotions: neutral (regural, typical state), disgust, funny state and interest state. The emotion analysis system was evaluated on 30 people (20 males, 10 females, age ranging from 24 to 42). All participants were presented with a close-up (field-of-view consisted mostly of the display) PowerPoint slideshow consisted of various photographs sorted on the type of emotion they were supposed to invoke (the playback time limit was 3 minutes for each of the emotional photo collection (same number of photos for each emotion)). The images were selected based on consulting with expert of human psychology. Up to 30 samples (pupil size, and x, y coordinates) are recorded during the one second.

In total there is more than 600 thousands of samples (Number of participants * Number of emotions * Minutes per emotion * Number of seconds * Number of per samples per second) which where divided in to training and testing data in the proportions respectively, 40% and 60%.

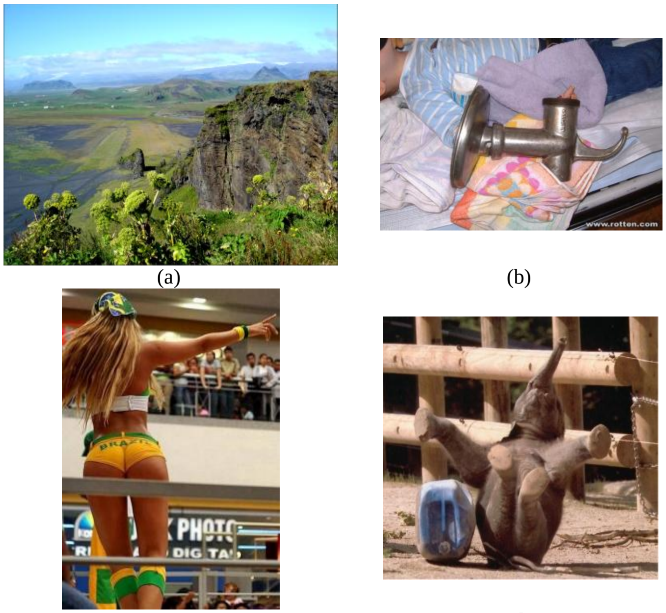

(a)
(b)
(c)
(d)

Fig. 3. The examples of visual stimulus which should invoke in the person four different emotions: a) neutral, b) disgust, c) interest and d) funny states.

After each set of emotional pictures there were 30 second pauses in the automated slideshow. During the experiment we have registered the size of the pupil, the coordinates of the center of the pupil in the video frame, as well as movement speed and acceleration. Figure 4 illustrates the fragment of the experimental analysis (6 people) on the size variation of eye pupil based on a current emotion (on the left) and size dispersion of eye pupil based on a current emotion (on the right).

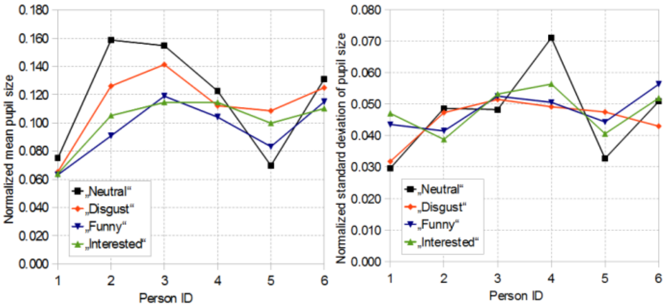


Fig. 4. The bar graph of relationship between average pupil size and the emotional reaction of the person (left), relationship between standard deviation of pupil size and emotional reaction (right).

The figure confirms the fact that the changes of the emotional stage can be recorded by observation of the small eye movements and the variation of the human eye pupil. Different emotional stimuli evoked different pupil sizes to different participant. For example, the pupil size of the first participant who was stimulated by the neutral visual stimulus is 28% bigger that pupil size that was measured when participant was stimulated with interested stimulus (see fig. 4 left). The right figure 4 shows the relation between variation of the pupil size in the time and the emotional stimulus.

The humans emotionally react differently to the different visual emotional stimulus. There can be, that one person feels the same emotions when he observes neutral stimulus and, another person, when he is stimulated with a funny stimulus. Every person interprets differently the emotional stimuli based on the life experience, emotional state at the beginning of the experiment. The correct answer cannot be given concerning on the automatically recognized emotion based only on the pupil size and variations, because pupil size can be affected by the general illumination, stress, physiological human properties and starting emotional status. Proposed emotion sensitive system uses not only information about the eye pupil size, but the classification features are enriched with the coordinates and motion speed.

Figures 5 and 6 illustrates an attention maps, which is computed based on the coordinate variation of the eye pupil center in the two dimensional space (measurement unit – normalized pixels). The attention map consists of green circle which radius depends on the time spent to observe certain part of the visual stimuli and the certain coordinates of the attention point. The attention map of the first participant is shown in the figure 4 and in the figure 6 the attention map of the second participant is shown. Overall registration period of the center of eye pupil is divided into four parts based on the shown emotional stimuli. The attention map shown in the figure 5 is computed from the data when participant were stimulated with a) neutral, b) disgust, c) funny and d) interested visual stimuli.

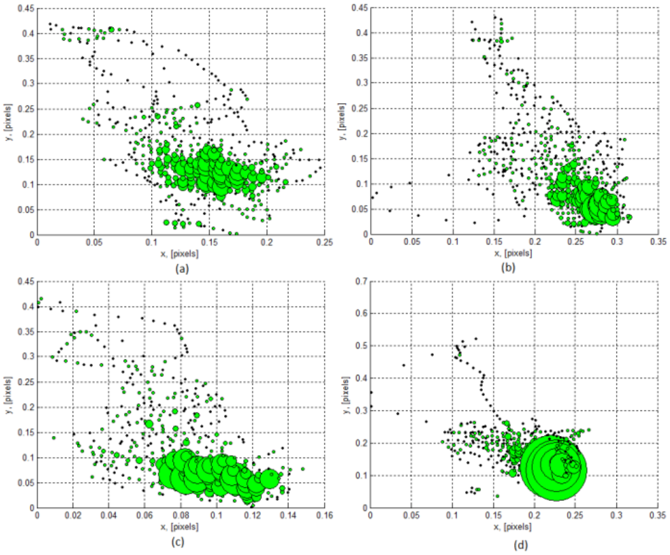


Fig. 5. The average variation of the attention point of the first experiment participant when he was stimulated with a) neutral, b) disgust, c) funny and d) interested stimuli

From the attention maps can be noticed, that the main attention concentration points are spread differently in two dimensional space due the different emotional stimulation. The attention points correlates with felt human emotion. The distribution of the attention points rely on the shown content of the visual information and the arrangement information on the screen. For example, all participants of the experimental investigation tend to avoid certain part of the disgust images.

The high distribution of the attention points has the data acquired during emotional stimulation using neutral images (see fig. 5a). The natural nature such as mountains and forest is captured in the recent visual stimuli; therefore, it can partly explain such big distribution of the attention points. The attention map of the second participant is slightly different in comparison with the attention maps of the first participant. When second participant was stimulated with the neutral stimulus, he was less concentrated then, when he was stimulated with funny stimulus (see fig. 6 a and c). It depends on the person's individuality, i.e., he finds that the funny stimulus is more interested (that shows the overlapping attention points in the figure 6c).

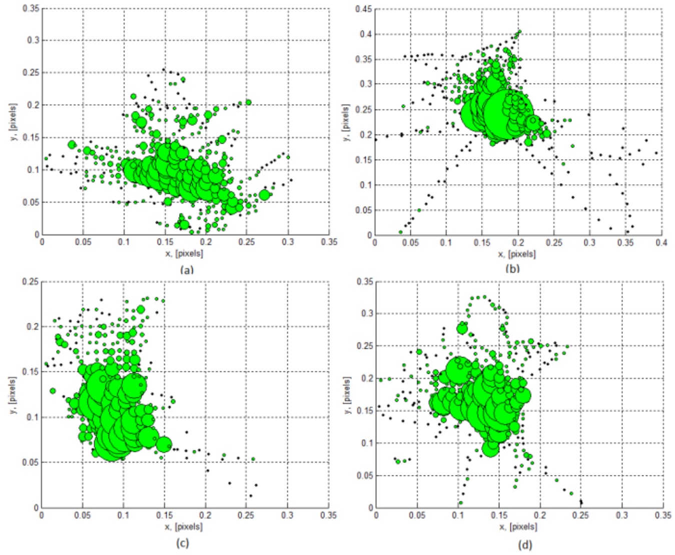


Fig. 6. An illustration of the average variation of the attention point of the second experiment participant when he was stimulated with a) neutral, b) disgust, c) funny and d) interested stimuli

Figure 7 illustrates the variation of the typical movement speed of eye pupil (a fragment of 6 persons) depending on the emotional stimuli. The motion speed of the eye pupil is measured in the pixels per second. The motion speed is presented on the vertical axis and the ID of the experiment participant is presented on the horizontal axis. The different color of the curve represents the different emotional stimuli. As it was with the variation of the pupil size, the motion speed depends on the person, on his starting emotional status, life experience and etc. For example, the motion speeds which were computed for the first participant differs in the relatively small range, i.e., up to 10%. While, the motion speeds of the sixth participant differs more than 46%. Such big difference may appear because of different cognitive capabilities, curiosity or usefulness of the presented information for the person (see fig. 7).

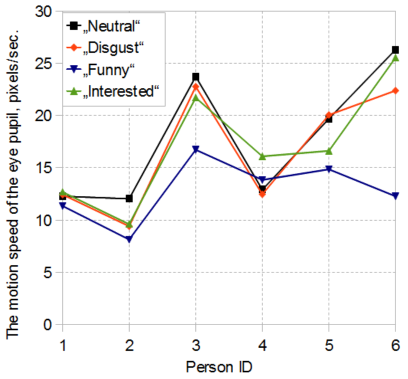
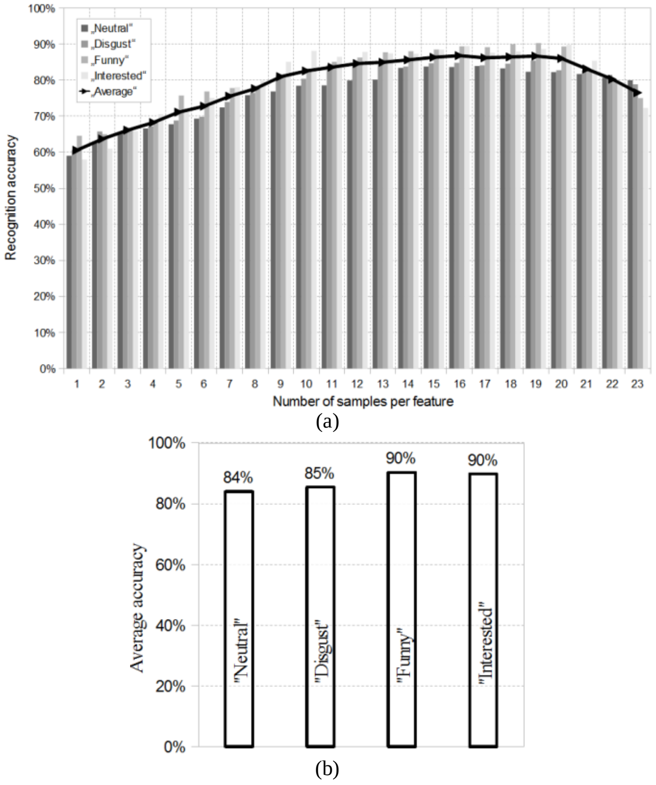


Fig. 7. The relationship between average speed of pupil motion and the emotional stimuli

Figure 8a illustrates the functional relationship between the number of feature samples and the recognition accuracy of different emotions. The overall best recognition accuracy (~90%) was achieved when we used 18 samples per feature. This means that the system can determine the emotion with a 2 second delay with approximately 10 % of deviation. The functional relationship shown in the figure 8a represents the accuracy graph of the one participant. The bar graph shown in the figure 8b represents the average recognition accuracy along the participants. Each bar represents the recognition rate for different emotion. Best recognition accuracy (90.27%) is reached when funny emotion is classified. Up to 16% of recognition fault is generated when system recognize the neutral emotion.



(a)

(b)

Fig. 8. The functional relationship between recognition accuracy and the number of samples per feature (a) and the bar graph of average accuracy (b)

IV. CONCLUSIONS AND FUTURE WORK

Experimental investigation has approved the fact, that the emotional state is individual and it depends on the persons cognitive perceptions. Although, it is possible to design a system based on the computational intelligence to recognize and detect certain emotional state of the human. The results of the experiments have shown that it is possible to detect the certain emotional state with up to 90% of recognition accuracy. Best recognition accuracy (90.27%) is reached when funny emotion is classified. Up to 16% of recognition fault is generated when system recognize the neutral emotion. The system can determine the emotion with a 2 second delay with approximately 10 % of deviation in average. Therefore, emotion recognition system uses 18 time samples per feature.

Future work will involve the application of remote eye tracking system that should allow recording more natural emotional responses to the visual and audio stimulation. There exist more than four emotional conditions; therefore, the multi-class problem will be solved using support vector machine and decision trees.

ACKNOWLEDGMENT

The works were funded by the Lithuanian Ministry of Education and Science according the project “Micro-sensors, micro-actuators and controllers from mechatronic systems” (GO-SMART, VP1-3.1-SMM-08-K-01-015).

REFERENCES

[1] Klin A, Schultz R & Cohen D (2000). Theory of mind in action: developmental perspectives on social neuroscience. In Understanding Other Minds: Perspectives from Developmental Neuroscience, ed. Baron-CohenS, Tager-FlusbergH & CohenD, 2nd edn, pp. 357–388. Oxford University Press, Oxford .

[2] M. Betke, J. Gips, and P. Fleming. The cameramouse: Visual tracking of body features to provide computer access for peoplewith severedisabilities. IEEE Transactions on Neural Systems and Rehabilitation Engineering, 10:1, pages 1–10, March 2002.

[3] M.A. Miglietta, G. Bochicchio, and T.M. Scalea. Computer-assisted communication for criticcally ill patients: a pilot study. The Journal of TRAUMA Injury, Infection, and Critical Care, Vol. 57, pages 488–493, September 2004.

[4] Arvid Kappas (Editor), Nicole C. Krämer (Editor) Face-to-Face Communication over the Internet: Emotions in a Web of Culture, Language, and Technology (Studies in Emotion and Social Interaction) [Hardcover] 316 pages Publisher: Cambridge University Press; 1 edition (July 25, 2011)

[5] Hess, Eckhard H. The tell-tale eye: How your eyes reveal hidden thoughts and emotions. Oxford, England: Van Nostrand Reinhold. (1975). xi 259 pp.

[6] Eckhard H. Hess and James M. Polt Pupil Size as Related to Interest Value of Visual Stimuli Science 5 August 1960: Vol. 132 no. 3423 pp. 349-350

[7] Amy D. Lykins, Marta Meana and Gretchen Kambe. Detection of Differential Viewing Patterns to Erotic and Non-Erotic Stimuli Using Eye-Tracking Methodology. Archives of Sexual Behavior. Volume 35, Number 5, 569-575

[8] Isaacowitz, Derek M. et al. Selective preference in visual fixation away from negative images in old age? An eye-tracking study. Psychology and Aging, Vol 21(1), Mar 2006, 40-48

[9] Proscevičius, Tomas; Raudonis, Vidas; Kairys, Artūras; Lipnickas, Arūnas; Simutis, Rimvydas. Autoassociative gaze tracking system based on artificial intelligence, Electronics and Electrical Engineering. ISSN 1392-1215. 2010, nr. 5(101), p. 67-72.

[10] Gengtao Zhou; Yongzhao Zhan; Jianming Zhang; , "Facial Expression Recognition Based on Selective Feature Extraction," *Sixth International Conference on Intelligent Systems Design and Applications, 2006. ISDA '06.*, vol.2, no., pp.412-417.

[11] Mihaela-Alexandra Puică, Adina-Magda Florea, Emotional Belief-Desire-Intention Agent Model: Previous Work and Proposed Architecture, International Journal of Advanced Research in Artificial Intelligence, Vol. 2, No. 2, 2013, pp. 1-8

[12] J. Grath, A domain – independent framework for modeling emotion, Journal of Cognitive Systems Research, 2004, vol. 4, No. 5, pp. 269-306

[13] I.B. Mauss, M.D. Robinson, Measures of emotion: A review, Cognition and Emotion, 2009, vol. 23, pp. 209-237

[14] C.J. Stam, Nonlinear dynamical analysis of EEG and MEG: Review of an emerging field, Clinical Neurophysiology, 2005, vol. 116, pp. 2266-2301

[15] O. Sourina, A Sourin, V. Kulish, EEG data driven animation and its application, In proceedings of international Conference Mirage, Springer, 2009, pp. 380-388

[16] Shylaja S, K N B. Murthy, S N. Nischith, Muthuraj R, Ajay S, Feed Forward Neural Network Based Eye Localization and Recognition Using Hough Transform, IJACSA,Vol. 2, No.3, 2011, pp. 104-109

[17] A. B. Watson, J. I. Yellott, A unified formula for light-adapted pupil size, Journal of Vision, 2012, 12(10):12, pp. 1-16

[18] P.N. Lopes, P. Solovey, R. Straus, Emotional intelligence, personality, and the perceived quality of social relationships, Personality and Individual Differences, 2003, Vol. 35., pp. 641-658

[19] Kohei Arai, Ronny Mardiyanto, Eye-Base Domestic Robot Allowing Patient to be Self-Services and Communications Remotely, International Journal of Advanced Research in Artificial Intelligence, Vol. 2, No. 2, 2013, pp. 29-33

[20] S.Nirmala Devi, Dr. S.P Rajagopalan, A study on Feature Selection Techniques in Bio-Informatics, International Journal of Advanced Computer Science and Applications, Vol. 2, No.1, 2011, pp. 138-144

Bootstrapping Domain Knowledge Exploration using Conceptual Mapping of Wikipedia

Mai Eldefrawi
Information System Department,
Faculty of computers and
information, Helwan University
Cairo, Egypt

Ahmed Sharaf eldin Ahmed
Information System Department,
Faculty of computers and
information, Helwan University
Cairo, Egypt

Adel Elsayed
Learning Systems International,
Leeds, UK, Formerly Research
Leader, M3C Lab, University of
Bolton, UK

***Abstract*—Wikipedia is one of the largest online encyclopedias that exist in a hypertext form. This nature prevents Wikipedia's potential to be fully discovered. Therefore the focus of this paper is on the role of domain knowledge in supporting the exploration of classical encyclopedic content, which in this case is Wikipedia. A main contribution provided by the author of this work is a methodology for identifying the nature, the form and the role of domain knowledge expressed in conceptual form. It's also a method of representation and analysis for describing the domain knowledge and for the extraction of the logical representation of a raw form of the domain knowledge. Such logical representation is of limited value in describing the real nature of domain knowledge. Hence we transform it into an adequate graphical representation, mostly of an arc-node form which is called conceptual representation.**

Keywords—Conceptual Mapping; Conceptual Representation; Domain knowledge; Wikipedia; Self-regulated learners.

I. INTRODUCTION

Domain knowledge exploration is one of key elements for learners specially when exploring open sources of knowledge. This exploration is not only a natural step for learners wishing to acquire new information related to their exploration interest, but it's a gate for adding new perspectives to their current knowledge from the wide range of available resources. The rapid technological age we are living has made it more demanding to develop new approaches and methodologies for exploring the knowledge available on the web. This requires a proper support for learners when exploring the web especially when they only have limited knowledge of a subject domain [11]. Learners who depend on online resources are usually self-regulated learners who need proper guidance. These online resources provide a rich and prosperous environment, but if not well managed learners can face a cognitive overload and distraction [13].

One of the online resources that are already available to help users get introductory information about specific domain is Wikipedia; Wikipedia is an online encyclopedia with around 30 million articles in 286 languages [22]. It's written collaboratively by volunteers around the word. Wikipedia has become very popular on the internet, with 365 million readers worldwide and ranked sixth among all worldwide websites. Wikipedia is available in a hypertext mode. The main disadvantage of Wikipedia that when navigating through its links each link takes the user to a different context far from the main idea being explored. As a result, there is no clear linkage between each topic and other external topics that could be of great importance to the reader. However the main power of Wikipedia is in containing a vast amount of concepts.

The open nature of Wikipedia raised concerns regarding the quality and consistent of information. These concerns led to an investigation conducted by Nature journal that showed that science articles have a very close accuracy rate to that of Encyclopedia Britannica [22]. Wikipedia articles are loosely organized; however any article usually starts with a short paragraph that summarizes the main features of an article with some definitions and hyperlinks. The scope of this paper considers this introductory paragraph of scientific domains within the analysis.

Wikipedia offers two features for navigating across article, the first feature is hyperlinks. Hyperlinks are merely an entry for other articles that exist in Wikipedia and mentioned by the name in the current article. The other feature is Categories [14], most users who are interested in finding the relations between topics and each other investigate the Wikipedia category tree.

Wikipedia category trees are arranged in the form of: main category>> a number of sub-categorical levels >> Pages. This hierarchy is the only arrangement for Wikipedia category tree, so an article is placed under another if according to the author of the article they are related. These articles may not necessarily be having a hierarchal structure according to their content modeling. However this categorization techniques isn't efficient enough for the following reasons.

- These categories are arranged by the author's point of view, this means that important articles may not be included in a category or vice verse.
- There is no clear criterion for including topics within categories and no provision over the referencing.
- Categories include entries under different contexts that are not explicitly stated, the name of an article is only mentioned with no elaboration [14].

Wikipedia categories trees are considered the most appropriate method for navigating a certain domain of knowledge within Wikipedia. Thus Wikipedia Categories are the focus of the analysis conducted in this paper.

Conceptual representation is an area that has been attracting researchers recently. As concepts form the base for human's cognition and graphical representation is usually more appealing for learners to read and understand, the use of graphical representation for concepts to support cognitive process has been the core of several studies. Novak defines concepts map as a graphical tools for representing knowledge. A concept map includes concepts contained in circles or boxes of some type. Relationships exist between pairs of concepts; these relationships are represented as a link connecting the related concepts. On the link the connecting word or phrase is placed which indicated the type of the relationship between the two concepts [21].

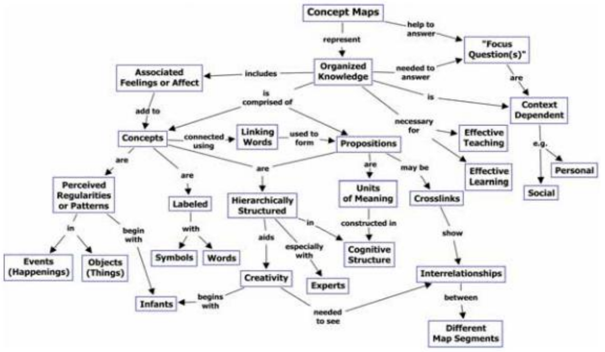


Fig. 1. A concept map showing the key features of concept maps. Concept maps tend to be read progressing from the top downward [21]

Our motivation is to benefit from Wikipedia by combining the advantage of it along with the advantage of conceptual representation. We are proving that using conceptual mapping for the analysis and exploration of a certain domain of knowledge within Wikipedia provide the user with the ability to navigate and explore through concepts, in a manner that couldn't be achieved by normal means. We are also comparing the findings with what the current navigational method of Wikipedia Provide, to show the potential that conceptual representation adds to the navigational and representation capabilities of Wikipedia.

II. RELATED WORK

The process of learning in a self-regulated manner is challenging as the learner can find himself overwhelmed with a flow of information in several directions and in many fields. The increased volume of knowledge, particularly when related to many fields makes the cognitive process more difficult to manage. As a result, the need for an adequate cognitive tool that allows users to easily construct their growing knowledge schema and navigate through large amount of available online resources become very demanding.

Knowledge based work demands ways of externally representing knowledge to complement the memory's functions. As a result, the idea of using concept maps in education was introduced. Conceptual mapping has a long history with education; early research in this field was conducted in [4]. They invented concept mapping technique to follow and understand changes in children's knowledge of science. Depending on the learning psychology of David Ausubel, that learning takes place by incorporating new concepts with existing concepts in the learner brain and generating relationship between them to make it easier for information to be memorized and retrieved when needed. This is known as "individual's cognitive structure". This idea was the basis for a lot of research especially in teaching, learning and assessment.

Empowering concepts mapping by incorporating it with the power of computers was the focus of several studies. For instance [5] suggested the use of advanced computer-based concept maps to help students in managing knowledge when they cope with the complexity of knowledge and Knowledge resources in many domains particularly in resource-based learning scenarios.

Experiments then were conducted to measure the effect of using Concept Maps as a knowledge construction tool. One of these experiments was conducted on a graduate online course [6]. In this course students were asked to construct a Concept Maps after reading four texts that are related in topic to one another. Students were asked to synthesize their understanding of the texts by representing graphically at least fifteen key knowledge objects. Overall observations show that students generally found the concept mapping activity useful but its full potential as a knowledge construction support tool was far from optimal. A similar experiment was conducted in [7], in which students were asked to study issues from their everyday life in a self-regulated way. Then the issues were organized and simulated by the cognitive approach with the use of computer-based conceptual mapping software.

The association of concepts to improve domain knowledge exploration is proposed in [11], in which a methodology that makes use of concepts association bank is introduced. This bank is used then to recommend new concepts within certain subject domain to knowledge explorers. The concepts association bank utilities the use of mind map to extracts concept associations from collective domain databases.

Domain knowledge exploration can also be supported through the enhancement of knowledge and knowledge resources themselves. With the availability of advanced computer-based concept-mapping tools, several researchers have discussed the potential of using digital concept maps to support spatial learning strategies and processes of individual knowledge management. In [13] the researchers not only showed that digital concept maps can support knowledge management and learning strategies, but he also showed that these tools can be used to represent content knowledge about a domain, as well as knowledge resources.

Representation and visualization of knowledge is another issue that faces self-regulated learners and knowledge explorers. One of the contributions of Concept Maps that they depend on the visual perspective of the receiver as the cognitive process becomes easier when concepts are plotted out in the form of map. However, this same property could result in a distortion as maps can get more and more complex due to the amount of information and concepts contained. For that reason visualization of concept maps and knowledge domain has become a challenge and a rich field for research.

A survey was conducted by [8] on visualization techniques of knowledge domains. The importance of this research that it doesn't only list previous work but it introduces bibliographic data set. This dataset includes articles from the citation analysis, bibliometrics, semantics, and visualization literatures. It applied different visualization techniques on the dataset and compared the results.

Also one of the interesting visualization of conceptual information space techniques is that discussed in [9]. It used the infolead technology to navigate a large number of web pages simultaneously in a 3D visual presentation. The research in this area is still open and evolving the main goal is how to make the interaction between the human and the machine more flexible and simulate the way the human brain works.

This is what led University of Wisconsin-Madison to adopt the project of CoMPASS (Concept Mapped Project-based Activity Scaffolding System) [1]. The CoMPASS is a hypertext system that uses two representations (concept maps and text) to enable multiple passes through the same material and to support inquiry and learning. The CoMPASS hypertext is used to help students generate ideas and learn about science concepts that will help them to solve their design challenges. Based on this system a study was made on middle school students [10]. The study showed that students who used the map version of the software their navigation was more focused and did better in the conceptual map and the essay test.

A. But how conceptual maps can be constructed?

Two main approaches have been suggested in [15] and [16] for producing conceptual maps. The first approach is a text-based approach [15]. As the name states text-based maps are generated closely related to certain piece of text. It is based on text-charting and Rhetorical Structure Theory (RST) and it's done in several stages. Starting with taking notes and summarizes of the text, Then a text charting is done on the text while considering the RST of the text to produce an initial map. The final stage results in the final concept map in which concepts and relationships between them is identified and plotted.

A method was suggested in [12] to facilitate the process of producing a text-based concept map; it is called Text-to-conceptual representation. The aim of this method is to facilitate the transformation of text-based information into a graphical arc-node conceptual representation. Traditionally the graphical conceptual representation was done directly while extracting concepts from a piece of text. So the conceptual map is being drawn while reading a piece of text. This required several re-correction and re-formulation and also makes it harder for users to trace back concepts in the map to the original text. Once the map is plotted there is no way to identify where each specific concept is extracted from. As the user depend only on his memory and brain to organized and extract the concepts without recording why and how he draw the map in this certain way.

Text-to- conceptual representation allowed for enriching the conceptual representation process. The conceptual outline extraction is done in a table of the following form.

TABLE I. TEXT TO CONCEPTUAL REPS CONVERSION – TEMPLATE [12].

Original text	*Conceptual Outline*	*Complementary Info*	*Media Assets*	*Supplementary Info*

- Original text: Sentence by sentence of the original text is placed in this column. The original text is placed here without change.
- Conceptual outline: In this section the conceptual outline is extracted and placed. It is extracted in the form of <Concept> predicate <Concept>.
- Complementary Info: This section contains any further information not expressed in the conceptual outline. This information is seen by the user as complementary and he can always go back to it for enhancing the understanding of the conceptual map.
- Media Assets: This column for any media (images, video) asset extracted from the text and adds to the understanding of it.
- Supplementary Info: any external information from sources other than the text in hand is added in this column.

As noticed from the above explanation of each column role in the creation of conceptual representation, the conceptual outline is recorded in this table. This recording allows the users to trace what they drew on the map back to the text. And also determine the origin of each concept from the text.

The other approach is domain-based approach; this approach is discussed in [16]. As our thoughts not always expressed in linguistic form; the domain-based approach is used in representing non-linguistic acquired knowledge. Although there is no specific definition for this approach, some main features that characterize this type of representation can be found in [16].

- It should cover the main attributes and features of certain domain of knowledge.
- It is not specifically related to certain cognitive schema.
- It is not specific to certain piece of text; it's a text free representation.
- The developed domain map should be of minimal representation and yet captures the main features of a domain.
- Graphical and visual icons can be used to describe the dynamics of certain domain of knowledge.
- The produced domain map should be incontestable to experts in the same domain.

The analysis of Wikipedia in this paper is done using text-based approach. However the resulting conceptual maps can be considered a domain map as explained below in the proposed method.

III. PROPOSED METHOD

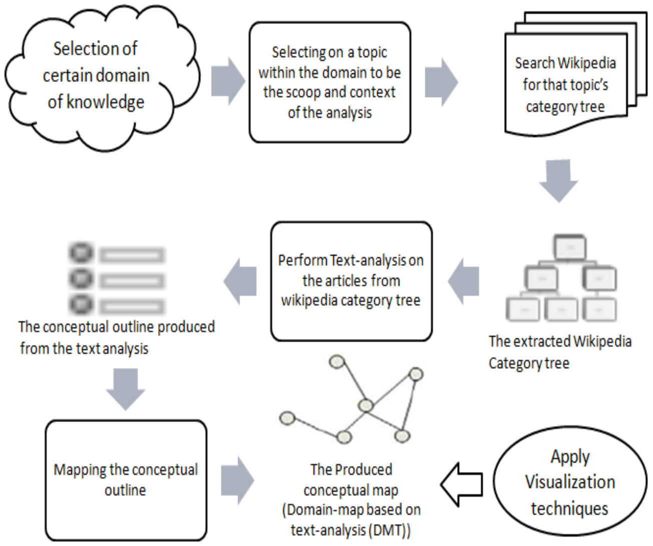

Selection of certain domain of knowledge
Selecting on a topic within the domain to be the scoop and context of the analysis
Search Wikipedia for that topic's category tree
The extracted Wikipedia Category tree
Perform Text-analysis on the articles from wikipedia category tree
The conceptual outline produced from the text analysis
Mapping the conceptual outline
The Produced conceptual map (Domain-map based on text-analysis (DMT))
Apply Visualization techniques

Fig. 2. Proposed method for analyzing Wikipedia domains

Although the use of conceptual maps to support self-regulated learners and to improve the navigation and exploration capabilities for learners has been addressed in several researches, the use of conceptual maps hasn't been applied before on this type of encyclopedic knowledge. Therefore it was necessarily to develop a new methodology to demonstrate how concept maps can add significantly to the representation and navigation process through encyclopedic domains of knowledge. The stages of the proposed methodology shown in Fig.2 the details of each stage are discussed below.

Selection of certain domain of knowledge

In this analysis the focus is scientific domains. Three domains were selected for the analysis, the criteria suggested for specifying domains is as follow: The first subject domain selected is a single subject single discipline, the second one is single subject multiple disciplines, the third one is a multiple subjects multiple disciplines.

Selecting a topic within a domain to be the scoop and context of the analysis

Within each domain a context should be stated clearly while navigating through Wikipedia. Such context makes it easier to judge on which concepts to include and which not in the analysis. This makes the domain of a manageable size. Without clear context so many concepts can be included which makes the analysis a daunting task. As when the chosen topic is subject group or principle subject, the navigated domain becomes so wide and harder to analyze and investigate.

Search Wikipedia for that topic's category tree

Each selected domain is searched in Wikipedia to extract the corresponding category tree of that domain. This extraction is simply done by finding the context related upper and lower categories of the selected topic along with any related pages within the categories' hierarchy. Categories or pages that are related to our context are included in the analysis that is done on the next phase. This relatedness is judged by the help of the experts in each domain.

Perform text analysis on the articles from Wikipedia category tree.

The text analysis performed on Wikipedia category tree articles is a text-based analysis done by Text-to- conceptual representation method mentioned above.

Mapping the conceptual outline

The produced conceptual outline from the previous stage is transformed into a conceptual map. The produced map is called Domain-map based on text-analysis (DMT).

Apply visualization techniques

It's widely known that when conceptual maps get larger in the number of displayed concepts and relationships, they began to look cluttered and more difficult to read. Another dimension of our research was to propose a method for the displaying of such large and nested maps. This leads us to the final stage of applying the visualization techniques. Visualization techniques are applied on the DMT map produced from the previous stage. These techniques are suggested to improve the visualization of the produced DMT map. In order to facilitate the exploration of such large and entangled maps an organization is of the map is proposed. The organization result in a 4 layers of the same map. These 4 layers end with the "Top cluster Level".

Layer4: is the bottom layer, this layer contains the DMT map as it is in its original form.

Layer3: in the 3rd layer, concepts are arranged according to their topical classifications. According to [18] topical classification is abstract structured spaces for arranging material spaces in which material or immaterial objects can get a location. Such immaterial objects can be concepts of certain domain or discipline or generally subjects of documents that are abstractly taken as information units. In this layer the relationships that join the concepts remain apparent.

Layer2: the resulting map from layer3 is displayed but the relationships are removed. This allows the user to concentrate on the concepts and see them without the interference of the relationships.

Layer1: the cluster layer, in this layer concepts are grouped into clusters with the header of each cluster is shown in the top of the class. Each cluster is about certain topic, the header is the main concept and the rest of the cluster is the concepts that are nesting from this main concept.

Domain Expert support

In all of the three explored domains experts supported the analysis in two main steps. Experts help essentially in reviewing the extracted Wikipedia category tree and the produced DMT map. In case of the category tree, evaluating the extracted map can cause the addition or removal of any concepts to the map according to their relatedness to the context in hand. In the case of the DMT map, they help in the comparison with the original category map.

It's been found according to the experts' provided support that the produced conceptual map can be considered a reasonable domain map. The analysis starts with text-based approach for the articles extracted from Wikipedia category trees and ends with producing a domain map that satisfies the domain map characteristics mentioned earlier. This solves an important issue regarding the creation of domain maps that require an expert to create them. So without the need for experts a domain map can be obtained through the text created by Wikipedia users, the reason for seeking a domain map is that these maps captures the main essential features of a domain without relying on a certain piece of text. Therefore the resulting conceptual map is called Domain-map based on text-analysis (DMT).

In each case we are assisted by different expert. In the first case the experts are teacher assistants and the authors of this work who work in the faculty of computers and information systems. In the second case the expert support is given by a domain map created in [23]. This map was created as teaching aid in the course of “introductory to signal processing”. In the final case the expert support is the category tree extracted from EduTechwiki. EduTechWiki is concerned with Educational Technology and related fields and hosted by TECFA - an educational technology research and teaching unit at University of Geneva. EdutechWiki is a resource for educational technology teaching and research. It also provide some (technical) tutorials for self-learners that or to be used in classes around the world [20]. EduTechWiki is organized in the same manner as Wikipedia, it include categories and sub-categories. As noted from the above experts are not necessarily human experts. The support can be given by a map or a reaserach conducted by some expert in certain domain. They act as experts by helping the analysist of this work to evaluate the results in comparison to artifacts in hands.

IV. Results and discussion

In this section the proposed method is applied on three cases, these three cases are discussed below.

A. Relational Database model case

1) Stage one: Selection of certain domain of knowledge

The domain selected in the first case is Database; there are two main reasons for choosing this domain. First this domain satisfies the predefined criteria in which database is considered a single subject, single principle domain

The other reason is that the authors of this work are professors and teacher assistant in the faculty of computers and information systems in the Information system department, and so considered experts in this area. We also consulted with another teacher assistant in the same department.

2) Stage two: Selecting on a topic within a domain to be the scoop and context of the analysis

For our analysis we selected relational database model [24] to be the context of this case.

3) Stage three: Search Wikipedia for that topic category tree

According to Wikipedia Relational Database model category tree is as shown in Fig.3.

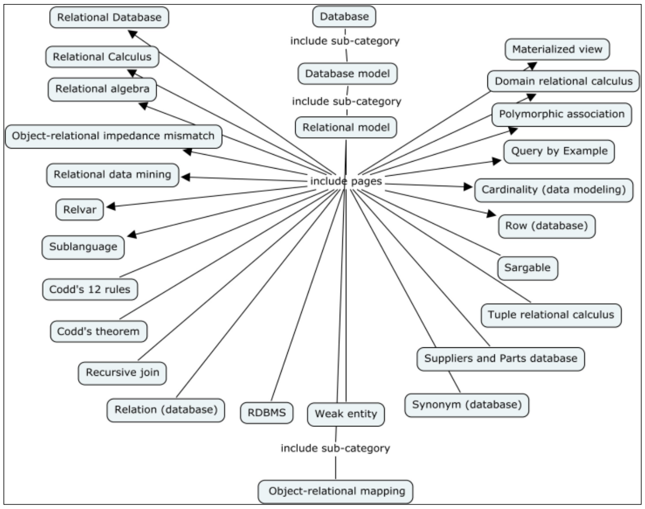
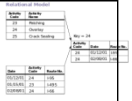
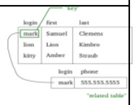


Fig. 3. Wikipedia Relational model Category tree

4) Stage four: Perform text analysis on the articles from the category tree

Below is sample of the text analysis done on the Relational Database model article.

TABLE 2. Text to conceptual representation of Relational database model article

Relational model Article analysis			
Original text	***Conceptual Outline***	***Complementary Info***	***Media Assets***
The relational model for database management is a database model based on first-order predicate logic, first formulated and proposed in 1969 by Edgar F. Codd.[1][2]	<Relational model> is a <DB model>		
In the relational model of a database, all data is represented in terms of tuples, grouped into relations	<Relational model> represent data in <Tuples> grouped into <relations>		
A database organized in terms of the relational model is a relational database.	<relational DB> organized according to <Relational model>		
The purpose of the relational model is to provide a declarative method for specifying data and queries	<relational model> is a <declarative method> for <data>		

	and <queries>		
users directly state what information the database contains and what information they want from it, and let the database management system software take care of describing data structures for storing the data and retrieval procedures for answering queries	<DBMS> describe <Data Structure > for <sorting> and <retrieval > of <data>	Users use DBMS to state what should be contained in the Database and what is the type of information they want to retrieve from it	
Most implementations of the relational model use the SQL data definition and query language	<relational model> implemented by <Sql DD> and <Query Language >		
However, SQL databases, including DB2, deviate from the relational model in many details; Codd fiercely argued against deviations that compromise the original principles		Another type of Sql Databases is DB2 that is considered a deviation from the relational model	

5) *Stage Five: Mapping the conceptual outline*

This process includes **transforming** the conceptual outline from the previous stage into a conceptual representation. The produced map from this stage is showed below for the relational model case, only a snap shot is shown below due to the size of the map.

Layer 4:

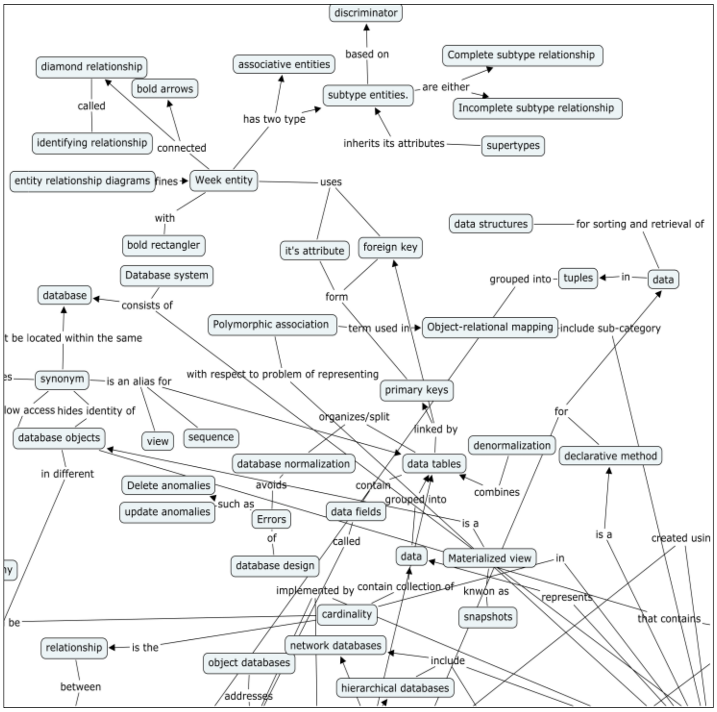


Fig. 4. Layer4 map, Conceptual representation of Relational model case (DMT map)

I. *Stage Six: Apply visualization techniques*

Layer 3:

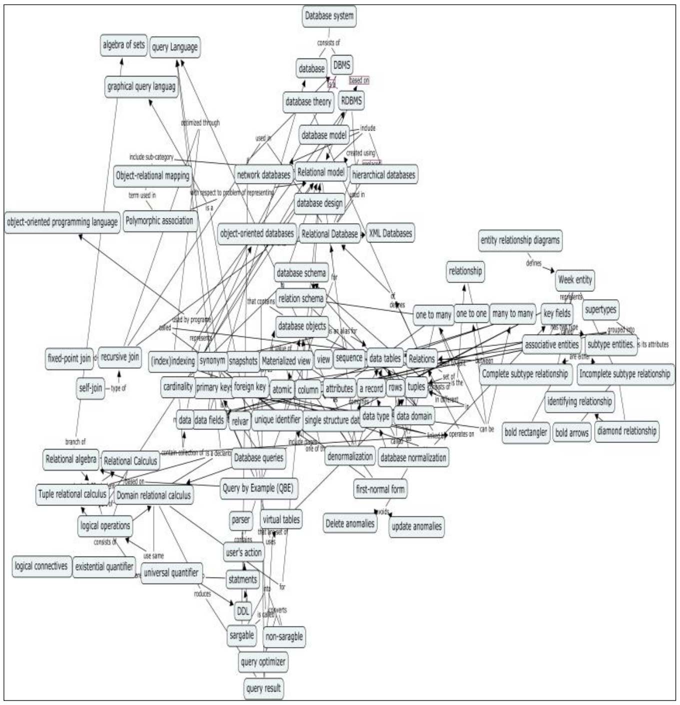


Fig. 5. Layer3 map, DMT map arranged according to concepts topical classifications

Layer 2:

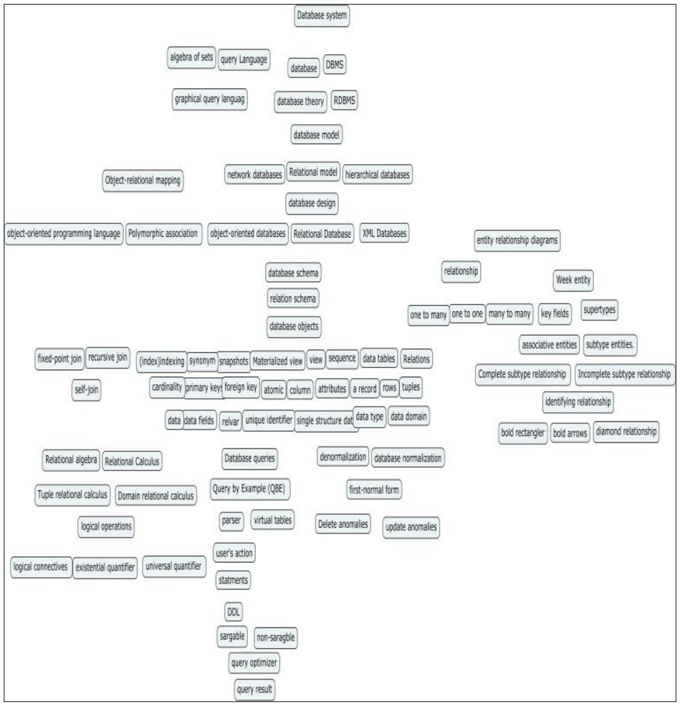


Fig. 6. Layer2 map, DMT map arranged according to concepts object classifications without relationships

Layer 1:

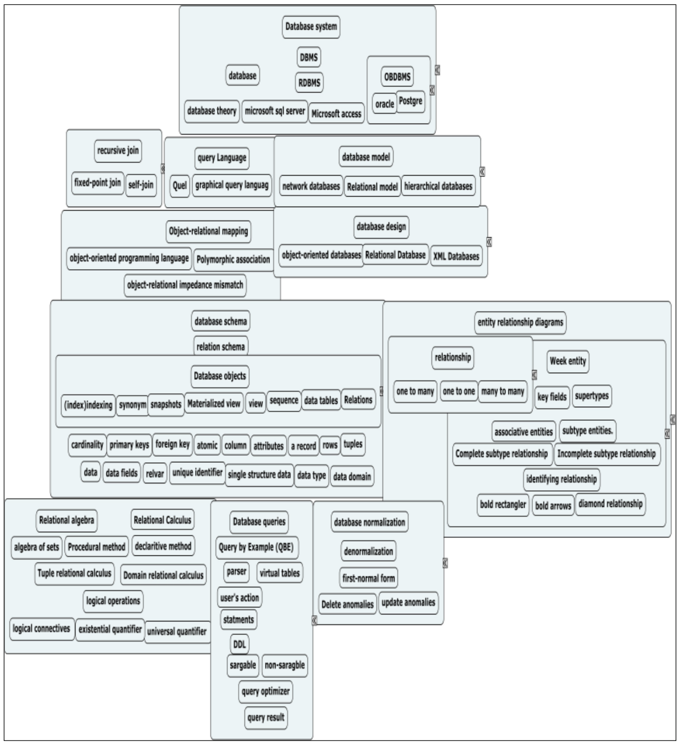


Fig. 7. Layer1 map, DMT map arranged into clusters of topics

Domain analysis:
Wikipedia category Tree:

Wikipedia Categories are arranged in a hierarchy which could lead to misleading information about what is related to what. For ex: in Fig.3, topic "RDBMS" is placed under "Relational model" as a subsidiary of it, while it's supposed to go under "DBMS" which is a software to implement and maintain Database. Note that "DBMS" along with "Database" form "Database system" which is in a higher level than "Relational model" in the Database hierarchy tree.

- An example of the misplacing of concepts under different categories. For ex: the topic "Week entity" exists while the "Entity Relationship diagram" that describes week entities and other types of entities isn't included.

- Some entries are placed under different contexts as in mentioning "Relational Algebra" and "Relational calculus"; these two topics form a theoretical foundation for query languages. A context where they are mentioned is usually when we are talking about Query Languages and the power of using them with "Relational Database". Other topics take us to different contexts like "Week entities" topic that is concerned with the establishment of a database schema.

- Fig.3 contains the topic "Relational Data mining", this topic describes a data mining technique applied on relational database; it must be place under "Data mining" category for those interested in data mining and its techniques. This is evidence that there is no clear criterion for including topics within categories.

The Resulting DMT maps: The maps produced by the proposed method shows significantly the difference between Wikipedia representation of domain knowledge and between the resulting conceptual representations. Even though our representation began by the analysis for articles from Wikipedia category tree and not from other sources, the results show wider views and interconnectivity between concepts within a domain.

The resulting cluster map in Fig.7 shows a variety of topics inside Relational model domain. For instance, Entity relationship diagram (ERD) that contains different types of entities including normal entities and week entities and the type of relationships between these entities are described clearly. Such information didn't appear in the original Category tree map, only mentioning week entity article as a page within the relational model category. Note that ERD is an essential concept in the creation and development of any relational database model. Fig.7 also contains the different types of Database objects, database models, query languages and concepts of normalization and denormalization. The different operations allowed by relational algebra and relational calculus are also placed in the map in Fig.7.

The Role of domain expert support: proving how well domain knowledge is represented is judged by the domain expert. Experts also help analyzing the included topics on both the category tree and the resulting DMT map. As explained above the resulting conceptual DMT maps provide a much more comprehensive view for the domain in hand. Also the relationships between the concepts are much easier to locate and understand.

In this case the evaluation process of Wikipedia category tree resulted in exluding some article like Relational data mining, Relver and other unrelated concepts.

B. *Simple Harmonic motion case*

The scope of the second case is Simple Harmonic Motion (SHM). SHM is a single subject, multiple disciplines. The concept of SHM is related to several areas; it's used in Physics, in mechanics, in engineering. It's also involved in so many dynamical systems for ex: "Pendulums", "mass-spring". And it's related to other motions like Circular motion. In the analysis of this case the expert support is a SHM map (expert map [23]) Fig.8.

after searching for SHM in Wikipedia, the extracted category tree is drawn as in Fig.9.

Then guided by the expert map the first category tree is expanded in which we searched Wikipedia for topics that exist in the expert map and add these articles to the first extracted category tree Fig.10. After applying all the six stages of our proposed method the resulting DMT map in the form of clusters (Layer 1) is then drawn in Fig.11.

Domain analysis:
Wikipedia category tree: For any user investigating SHM topic, the first category tree Fig.9 is what he will find in Wikipedia. This shows the difference between a map truly describing the domain and what users who don't have enough knowledge will find.

In the expanded category tree in Fig.11 it included important articles like, oscillation, waveform, amplitude, damping, sine wave, circular motion, spring pendulum and pendulum. This shows the deficiency of navigating Wikipedia without previous knowledge in a domain. Without the guide map such articles that are according to the guide map form a very important aspects of SHM topic wouldn't been reached. For users who are reading for the first time about SHM, these topics will look like any other topic on mechanics category that contains over 200 categories, sub-categories and hundreds of pages under these categories.

- Wikipedia categorization trees have ad hoc organization of topics insufficient to fully and satisfactory characterize a domain. For ex: topic like "Oscillations" is categorized in the same level with classical mechanics, and treated as a topic of higher level than SHM. While in fact oscillation is one of the behavioral features that describe SHM it is part of it.

- Concepts that are supposed to be linked to each other appear here as separate entities not related in the Wikipedia category tree. For ex: dynamical systems that contains systems such as "Mass-spring", "Pendulums", and "spring pendulums".

- When searching for concepts related to engineering and SHM, the only results are instruments related to the control part of engineering and not the motion.Even when searching for the two concepts together "Pendulum+ dynamical systems" although the pendulum concept exist in "dynamical system" it didn't show in the results and didn't reflect in the category tree of them together.

- Also the steady state and transient states of system behavioral features exist under different categories far from SHM

The Resulting DMT maps: This classification shows the domain of SHM in a more detailed and clearer representation. It specifies the natural phenomenon's affecting any system under motion like (frequency, gravity, oscillation), different types of dynamical systems that are subject to motion like (springs, pendulums, oscillators), different types of periodic motion like (cicular motion, SHM, molecular motion), damping and their types that affects the motion of dynamical systems and the states these systems can be under.

The resulting conceptual maps don't simply list the concepts, but they also arrange them and specify the type or relationships between them. This shows the importance of the conceptual representation in detecting concepts that are highly related to the domain and yet are not included or included in ambiguous way.

The role of domain expert support: The first category map in Fig.9 shows that normal users will only find articles about simple harmonic motion, classical mechanics and pendulums. It is almost impossible for users with no knowledge about the domain to find the rest of the topics that acquired by using the expert map in Fig.8.

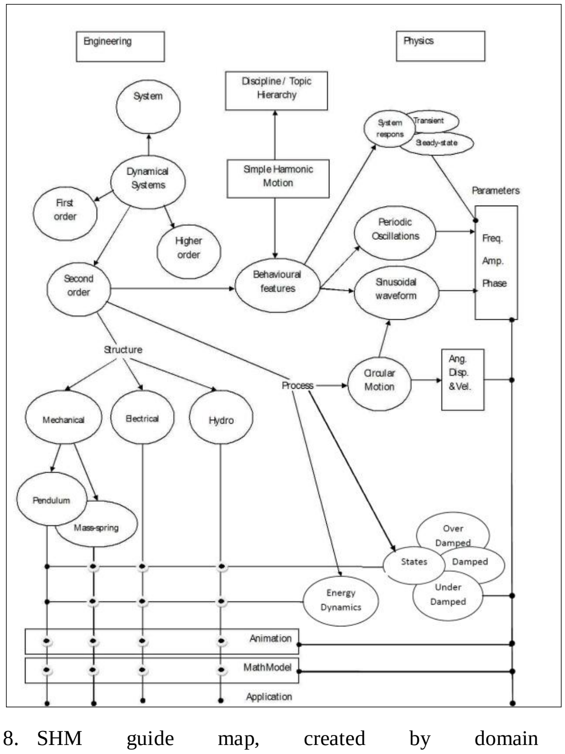


Fig. 8. SHM guide map, created by domain expert

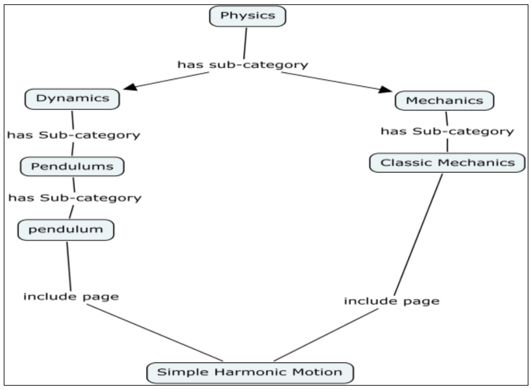


Fig. 9. Wikipedia SHM Category tree

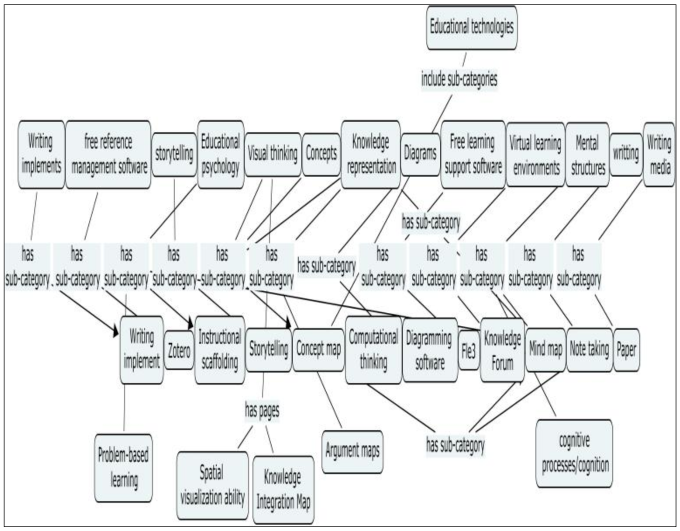


Fig. 10. SHM Category based on SHM Guide map in whole of Wikipedia

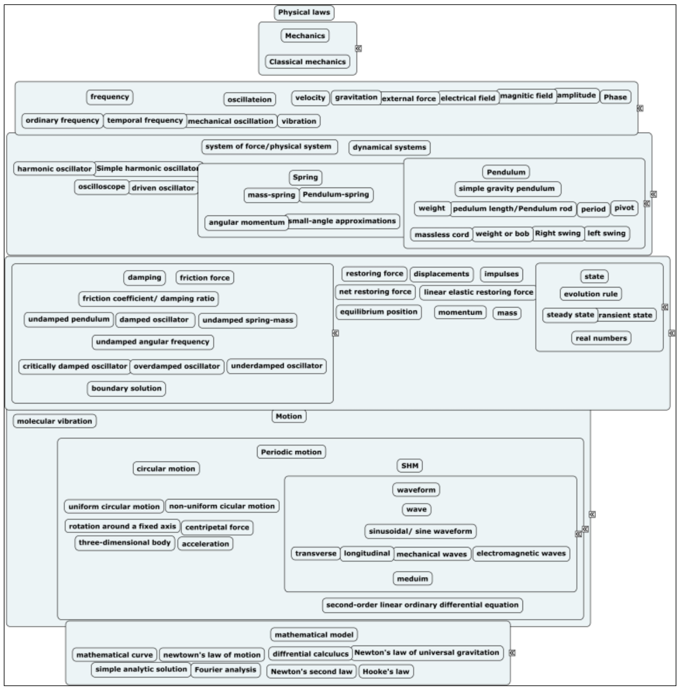


Fig. 11. Layer1 map, DMT map arranged into clusters of topics

C. Cognitive support tools

The third case is Cognitive support topic. Following same selection criteria as done on the previous 2 cases, this case is a multiple subjects, multiple disciplines. The selected topic within cognitive support domain is cognitive support tools.

When searching for this topic 'cognitive tools' in Wikipedia, this topic wasn't found in Wikipedia, not even the topic 'cognitive support'. In this case, the expert support that will guide the search in Wikipedia is EduTechWiki [19]. Guided by category tree of EdutectWiki map in Fig.12 a search is conducted in Wikipedia for finding topics that matches those in Fig.12. The resulting Wikipedia category tree for cognitive support tools is drawn in Fig.13. Although the main concept 'cognitive tool' doesn't exist in Wikipedia, some of the articles in EduTechWiki category tree exist in Wikipedia but under different categories. After applying all the six stages of the proposed method the resulting DMT map in the form of clusters (Layer 1) is drawn in Fig.14.

Domain analysis:
Wikipedia category tree: Same issues regarding the organization of Wikipedia categories can be seen in this case too. Fig.19 shows that concepts that are related to cognitive tools appear here to be scattered all over Wikipedia under different categories. Only few articles seem to have some sort of a connection as Note taking, Mind map and Concept map, Instructional scaffolding and Note taking.

The Resulting DMT maps: Although the concepts of Cognitive tools and cognitive support are not explicitly mentioned in the articles extracted from Wikipedia but after clustering these concepts start to surface. Cluster like Cognitive skills include concepts that describes cognition and skills acquired by the brain. Also the Cognitive support cluster shows type of support that can be given to learners in order to support the cognitive skills. It also shows the type of mental processes and functions that human brain is capable of.

The cognitive tools cluster is a group of the cognitive tools that can support the cognitive process performed by the brain. Some of these tools use visualization techniques like this approach in hand 'conceptual map', so another cluster appears that include some of the visualization features used in the cognitive process.

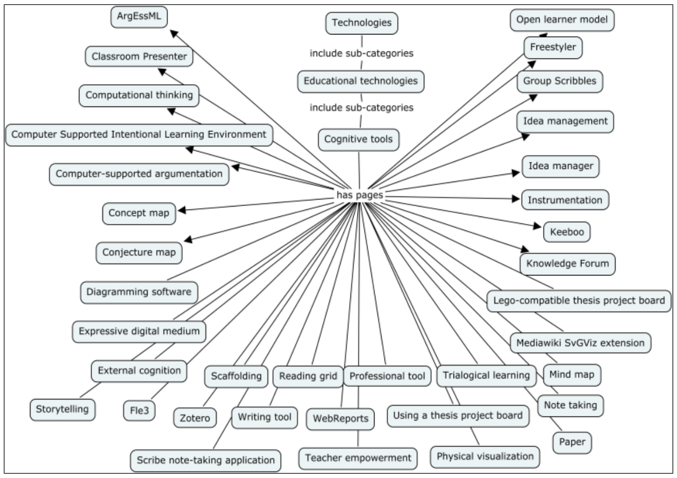


Fig. 12. EduTechWiki cognitive tools category tree

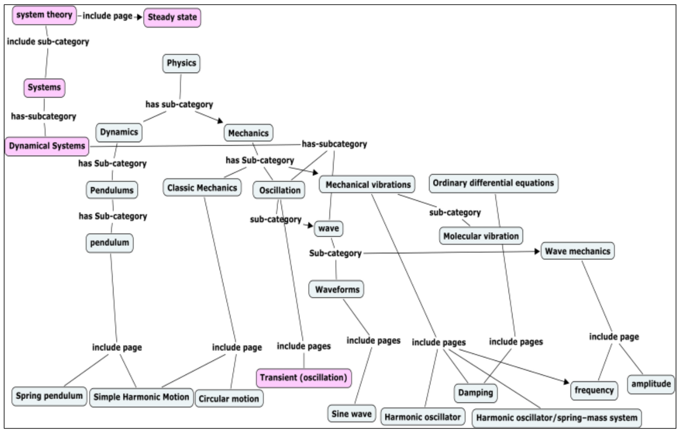


Fig. 13. Wikipedia category tree of Cognitive support tools

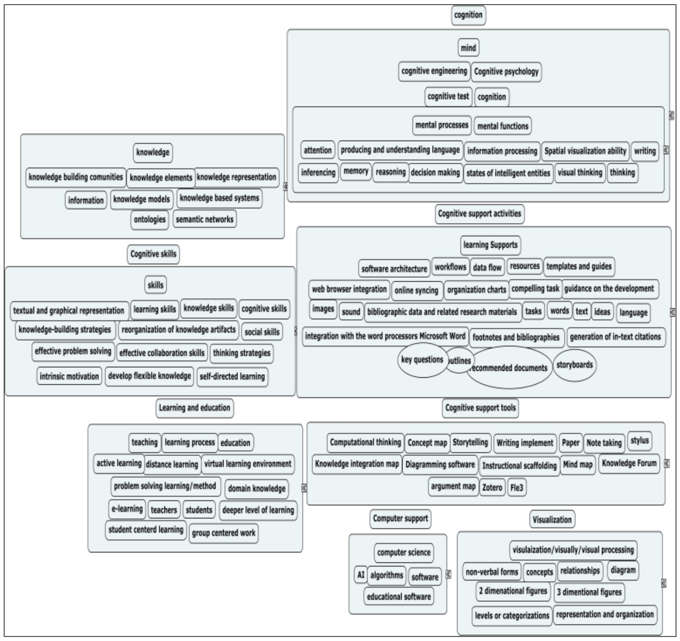


Fig. 14. Layer1 map, DMT map arranged into clusters of topics

Also a new tool like argument map and Knowledge integration map weren't mentioned in the EduTechwiki category tree. But it appeared in the conceptual analysis of the articles as tools that support the cognitive process.

The Role of domain Expert: This topic in particular is very important in this analysis, as the topic in hand doesn't exist in Wikipedia as a topic. Obviously this is a dead end for any knowledge explorer who would simply stop searching. Although cognitive tools are not explicitly stated in Wikipedia but several tools are mentioned under different categories. Guided by the Edutechwiki category tree the resulting DMT map shows the topic in a clear way with concepts didn't appear in the original category tree.

V. CONCLUSION AND FUTURE WORK

In this paper a new methodology is introduced for representing and navigating online encyclopedic knowledge using conceptual representation. This methodology combined the two approaches of creating conceptual maps, starting with text-analysis of Wikipedia articles to end up with a comprehensive domain map of the analyzed domains of knowledge. The support of experts is essential in evaluating the produced results and they specially helped in comparing between the initial Wikipedia category tree map and the resulting DMT map.

The resulting DMT maps showed unarguable results in describing domains of knowledge in a way that Wikipedia couldn't achieve before. This leads to better understanding of a domain and a better navigation across different concepts within the same domain. Users initially get lost by the amount of articles within each category those are not necessarily related to the domain and limited by the navigational approaches offered by Wikipedia which are hyperlinks and categories. By representing a domain in a conceptual map, the whole domain is plotted in a single map that a user can zoom in and out and choose which concepts to view and which not.

This research can be extended in the future by placing the produced conceptual DMT maps as a front end for Wikipedia users, these maps won't replace the original text but they will act as a navigational gate for the domain underlying these maps. Users can choose at any point to navigate further down and read the details that exist in each article. Users will also be able to add concepts to the map and edit it the way Wikipedia offers now, so users themselves will be able to see the domain map and decide what is missing or what can be added.

The focus of this paper was on scientific domains, this research can be experimented on other domains to measure how effective the proposed method can be with other domains. The proposed method can also be used to classify and determine the nature of wikipedia different domains.

REFERENCES

[1] CoMPASS, "Concept mapped project-based activity scaffolding system", http://www.compassproject.net/info/index.html, Wisconsin Center for Education Research. 2008. Web. 24 May 2011,

[2] J. A. Hampton, H. E. Moss, "Concepts and meaning: Introduction to the special issue on conceptual representation". *Language and Cognitive Processes*, 18(5-6), pp. 505-512, 2003

[3] A. El Sayed," Conceptual Representations a content modeling perspective", unpublished, 2011.

[4] J.D. Novak, D.B .Gowin, “Learning how to learn”, Cambridge: Cambridge University Press, 1984

[5] S.O.Tergan, “Concept Maps for Managing Individual Knowledge”, Proceeding of the first joint meeting of theEARLI, pp. 229- 238, 2004

[6] J.Basque, B.Pudelko, “Using a concept mapping software as a knowledge construction tool in a graduate online course”, Proceedings of World Conference on Educational Multimedia, Hypermedia and Telecommunications,
pp. 2268-2274, 2003

[7] W. Gräber, A. Neumann, S.O. Tergan , “Mind Mapping for Fostering Self-Regulated Resource-Based Learning in Science Classes”, International Workshop on Visual Artifacts for the Organization of Information and Knowledge, May 13th - 14th, 2004.

[8] K.Börner, C.Chen, K. W. Boyack, , “Visualizing knowledge domains”, Annual Review of Information Science and Technology , Vol. 37 , Nr. 1 Wiley, S. 179-255, 2003.

[9] K.Toru , K.Atsushi , T-U.Kaku , M.Tetsuyuki , “A Proposed Net-space Service Using InfoLead Cruising Navigation Technology”, IEEE Computer Society. pp. 31-32, 2002.

[10] S.Puntambekar, A.Stylianou, R.Hübscher,” Improving navigation and learning in hypertext environments with navigable concept maps”. Human Computer Interaction, 18 (4), pp.395-426, 2003.

[11] C-C. Liu, S-H.F.Chiang, C-Y.Chou, S.Y. Chen, “Knowledge exploration with concept association techniques”, Online Information Review, Vol. 34 Iss: 5, pp.786 - 805, 2010.

[12] A.Elsayed. “A text analysis methodology for text-based conceptual modeling of knowledge domains”, Technical Report, Learning Systems International, Leeds, UK, 2012.

[13] S.O.Tergan, “The Use of Digital Concept Maps as Cognitive Tools for Managing Knowledge and Knowledge Resources”, AAAI, symposia, SS-05-06, pp.73-76, 2005.

[14] Wikipedia,”Categories, lists, and navigation templates”, http://en.wikipedia.org/wiki/Wikipedia:Categories,_lists,_and_navigation_templates, 2001.web. 6 Feb. 2013.

[15] H.Roger,” Conceptual Representations and the Dimensions of Textual Content” , M3C International Workshop taking place at (ICALT), University of Bolton, UK, July 2009

[16] A.Elsayed ,"Relations for Graphical Conceptual Representations (Presented Conference Paper style)", M3C International Workshop taking place at (ICALT), University of Bolton, UK, July 2009

[17] Higher Education Statistics Agency,” JACS3 Classification”, http://www.hesa.ac.uk/content/view/1787/281/, 1993. Web. 5 Sept. 2012.

[18] De. Antonella , M. Dario, M.Alberto, “Subject Classifications in the Scientific and Overall Digital World”, High Energy Physics Libraries Webzine, Draft, 2001.

[19] EduTechWiki,”Cognitive tools”, http://edutechwiki.unige.ch/en/Cognitive_tool
, 2001. Web.5 Mar. 2013.

[20] EduTechWiki, “Main Page”, http://edutechwiki.unige.ch/en/Main_Page
, 2001. Web.5 Mar. 2013.

[21] J.Novak, C.J. Alberto, “The Theory Underlying Concept Maps and How to Construct and Use Them “, Technical Report IHMC CmapTools 2006-01.

[22] Wikipedia, “Wikipedia”, http://en.wikipedia.org/wiki/Wikipedia
, 2001.web. 6 May 2013,

[23] A.Elsayed, “A domain representation approach for the use of encyclopaedic knowledge”, Technical Note, M3C Lab, University of Bolton, UK, 2009.

[24] School of Mathematical & Computer Sciences, “F21DF1 Database and Information Systems”, http://www.macs.hw.ac.uk/~trinder/DbInfSystems/l4RelModel2up.pdf, 2010.web. Nov.2010.

Dynamic Evaluation and Visualisation of the Quality and Reliability of Sensor Data Sources

Ritaban Dutta, Claire D'Este, Ahsan Morshed, Daniel Smith
Intelligent Sensing and Systems Laboratory
CSIRO, Hobart, Australia

Aruneema Das, Jagannath Aryal
University of Tasmania
Hobart, Australia

***Abstract*—Before using remote data sources, or those from external organisations, it is important to establish if the source is fit for purpose. We have developed an approach to automatic sensor data annotation and visualisation that evaluates overall sensor network performance and data quality. The CSIRO's South Esk hydrological sensor web combines data related to water management from five different organisations, which provides a suitable platform to explore the issues of reliability and uncertainty. An environmental gridded surface is generated based on the observations and evaluations of quality and reliability of the sensor node provider.**

Keywords—Dynamic Visualisation; Sensor Web; Weather Station; Time Series Catalogue; Interpolated 3D Surface.

I. Introduction

Given the high cost of hardware, technical overhead, and significant maintenance required by environmental sensor networks there has been a shift towards sharing of data to distribute the load on organisations. Sensor data must be provided through means that promote re-use [1]. Standards such as Sensor Web Enablement from the Open Geospatial Consortium (OGC) [2] encourage sensor data interoperability. Web and cloud services assist with distributed data storage and public accessibility. However, these approaches do not provide assurances of the reliability and data quality of a sensor web.

Due to an improvement in the transparency of recent sensor network and communication technologies, the uncertainty associated with environmental sensor webs is becoming increasingly evident. This uncertainty is commonly associated with the limited availability of data (spatially and temporally) and/or the poor quality of the available data. The sensors are often deployed unattended in harsh operational environments. The external causes of sensor uncertainty include their operation under extreme conditions, calibration drift, and bio-fouling. In addition, sensor nodes are subject to communication, software, electronic, and battery failures.

Next generation environmental monitoring, natural resource management and related forecast-based decision support systems are becoming increasingly dependent on web-based data integration of large scale sensor networks. This integration requires different forms of pre-processing, including accumulation and harmonization. Automating the verification of the quality of the individual sources is essential to build trust in the users of these systems. A tool to analyse and visualise the uncertainty of data sources is required to determine whether the sources are complementary for the purpose of integration [3].

We have designed and developed an adaptive tool for analysis and visualisation of the South Esk hydrological sensor web around a general theme of geographical location information. Near real-time analysis of sensors was performed in relation to the quality and availability of its data. Data cleansing and imputation techniques were also applied according to the weather station manufacturer's sensor specifications and attribute value range validations using well defined hydrological knowledgebase. Finally this study estimated and visualized a dynamic 3D surface map of the South Esk region based upon available environment data.

II. Related Work

Automated assessment of sensor data has been explored in the marine domain to provide quality flags for data consumers [4].

There has been some attempt to quantify the reliability of a wireless sensor network. Most often the reliability analysis is performed with a probability graph [5] and include measures of factors such as fault diagnoses, analysis and recovery. Purohit et al. [6] modeled the hardware and software modules of a wireless sensor network as a series-parallel structure using a reliability block diagram approach. The problem can also be decomposed into sub-problems using a physical model to align with the dynamic nature of a sensor network [7].

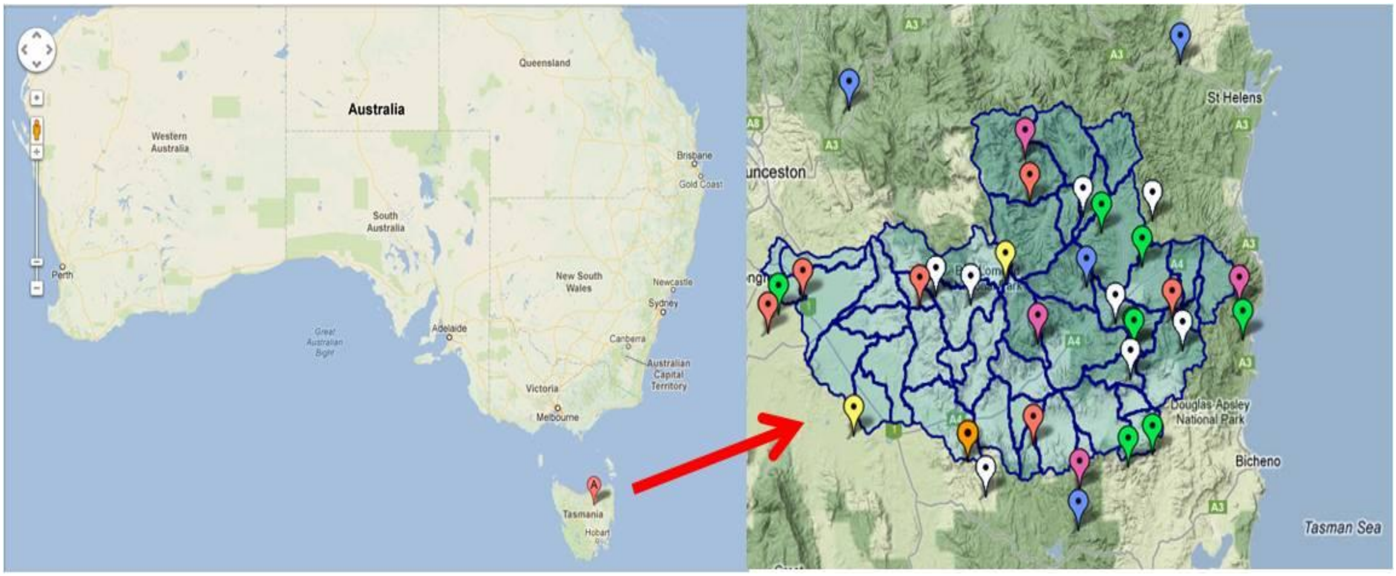


Fig. 1. The Google Maps™ pane presents a federated view of near real-time sensor data from the different sensor networks operating in the South Esk catchment (different colours correspond to sensors from different agencies).

III. The South Esk Hydrological Sensor Web

A. The Sensor Network

The Sensor Web is an advanced spatial data infrastructure in which different sensor assets can be combined to create a sensing macro-instrument. This macro-instrument can be instantiated in many ways to achieve multi-modal observations across different spatial and temporal scales.

CSIRO is investigating how emerging standards and specifications for Sensor Web Enablement can be applied to the hydrological domain. To this end, CSIRO has implemented a Hydrological Sensor Web in the South Esk river catchment in NE Tasmania (Figure 1). The South Esk river catchment was chosen because of its size (3350 square kilometers, large enough to show up differences in catchment response to rainfall events), spatial variability in climate (an approximate 800mm range in average annual rainfall across the catchment), variable nature of seasonal flow, and relatively high level of instrumentation.

This is made possible by re-publishing near real-time sensor data provided by the Bureau of Meteorology (BoM), Hydro Tasmania, Tasmania Department of Primary Industries, Parks, Wildlife and Environment (DPIPWE), Forestry Tasmania and CSIRO via a standard web service interface (Sensor Observation Service) developed by the Open Geospatial Consortium (OGC). The specific SOS implementation was developed by the 52° North Initiative. Enhanced situational awareness of the catchment is gained by exposing sensor data via standard web service interfaces [8].

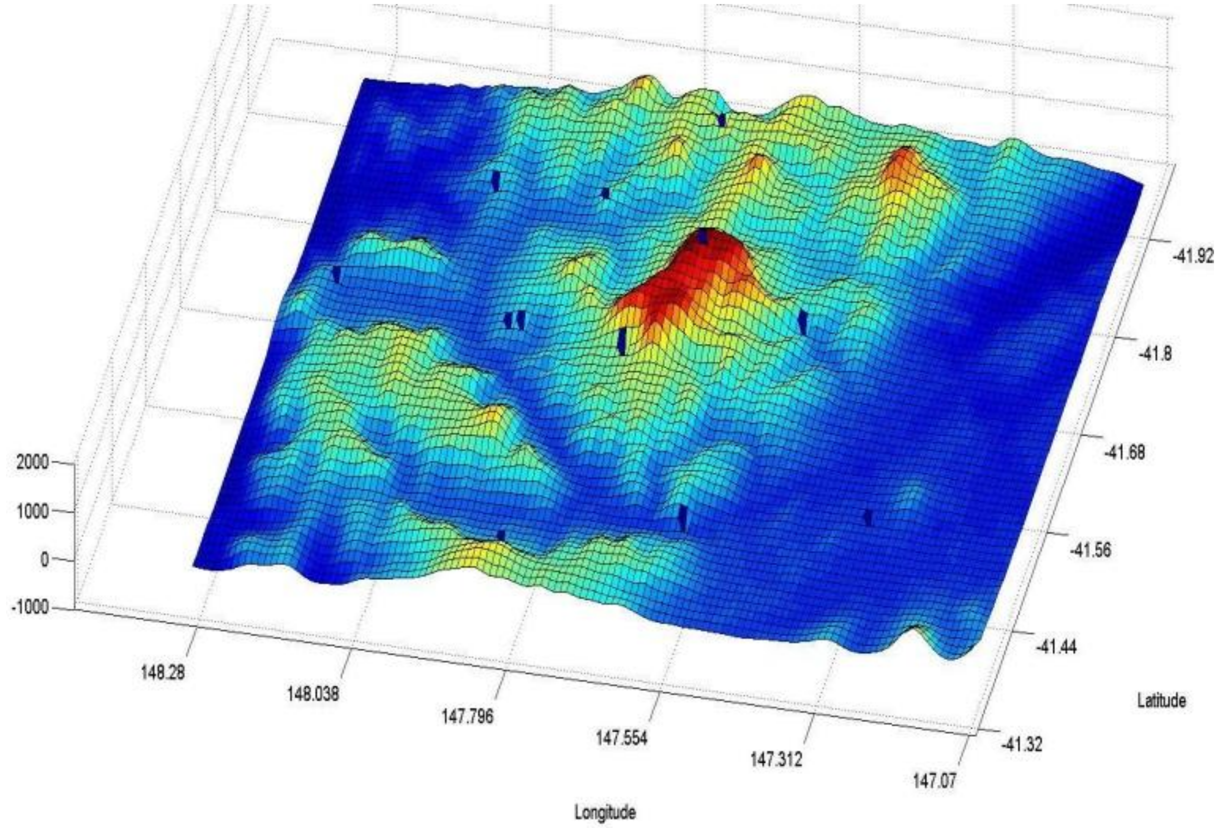


Fig. 2. Three dimensional map based on South Esk catchment topography data. Distributed sensor nodes locations are represented in dark blue.

B. *Coordinate System*

The South Esk sensor web covers an approximately 95 km × 220 km rectangular region. It covers a latitude range between 40.5°S and 43.5°S and a longitude range between 145°E and 148.5°E.

Part of the study involved developing a South Esk Data Service Tool, which generates visualizations such as Figure 2. In this figure, the South Esk catchment is depicted as a 3D surface based on patched elevation data. The whole region was mapped as a gridded rectangle where each cell represents a 5km × 5 km region. The physical locations of the weather station sensor nodes are visible as blue marks on the 3D surface.

C. *Weather Stations*

A small number of RIMCO 7499 tipping bucket rain gauges are operated by BoM in their weather stations.

Vaisala automatic weather stations made up the majority of the sensor network.

TABLE I. METEOROLOGICAL SENSORS IN THE SOUTH ESK SENSOR WEB

Sensor	Phenomenon	Method	Data Validity Range
Vaisala WMT52	Wind speed	Ultrasound	0 – 60 metres per second
Vaisala WM30	Wind speed	Ultrasound	0.5 – 60 metres per second
Vaisala RAINCAP	Precipitation	Acoustic impact of rain drops	0 – 200 millimetres per hour
RIMCO 7499	Precipitation	Tipping bucket	0.2 millimetre bucket
Vaisala HMP45A	Humidity	Glass substrate	-40 - +60 degrees Celcius
Vaisala HMP45A	Temperature	Passive thermocouple	0.8 – 100% relative humidity
RIMCO EQ08 & EQ08E	Total Global Radiation	Pyrano-Albedometer	0 – 2000 watts per square metre
RIMCO EP 16 & EP 16E	Upward/Downward Solar Radiation	Dual head (as above)	0 – 1400 watts per square metre
RIMCO Middleton SK08	Solar Radiation	Pyranometer plane surface	0 – 20000 watts per square metre

The Vaisala weather transmitter (WXT520) measures phenomena including barometric

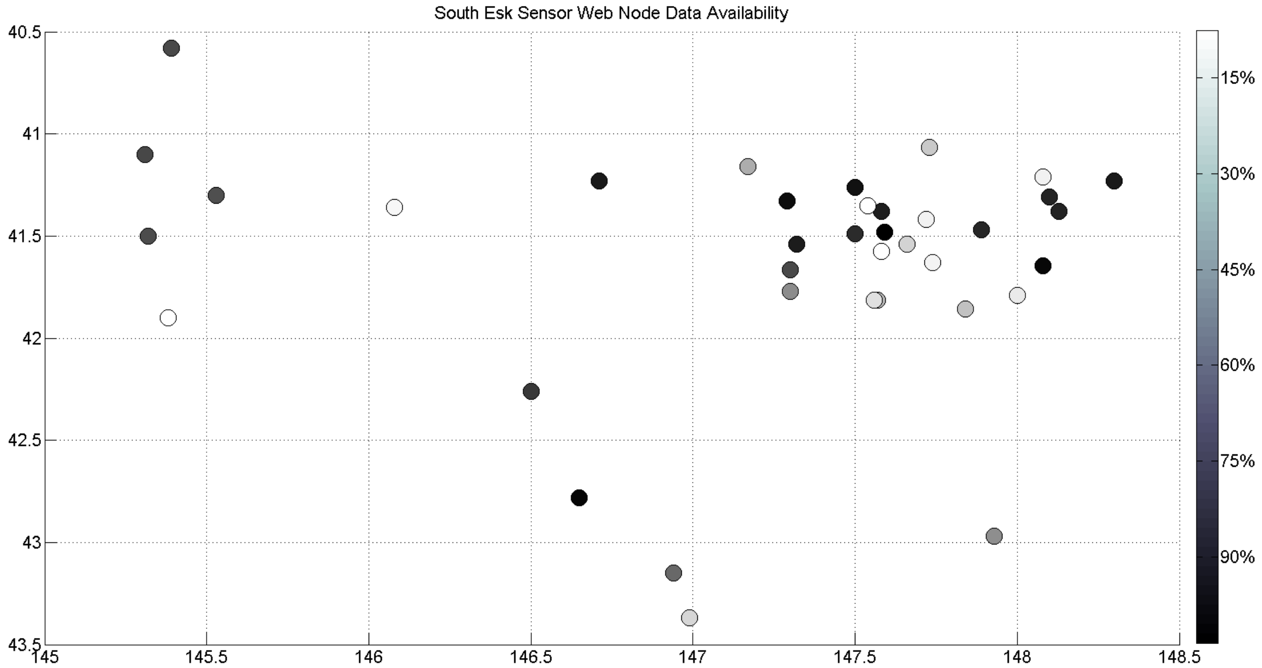


Fig. 3. Node based data availability for the South Esk sensor web. Darker node represents higher data availability with low uncertainty.

pressure, humidity, precipitation, temperature and wind speed.

Table 1 lists the sensors available in the South Esk including their valid data ranges. The valid measuring ranges were used later for dynamic data filtering purposes.

IV. Sensor Web Visualisation

A facility to query near real-time status updates from individual nodes in the sensor network was developed. The resulting system, the Timeseries Catalog, uses the available web services for this purpose.

For the visualisation study, we focused on five phenomena:

- temperature
- relative humidity
- rainfall
- solar radiation
- wind speed

The observations of these phenomena were acquired from the 40 sensor nodes in the South Esk hydrological sensor web. A data service was developed to search, extract and download time series using the Timeseries Catalog hypertext transfer protocol [9].

A. South Esk Data Service Tool

Figure 4 shows a sample from the data visualisation tool developed for this study, the South Esk Data Service Tool graphical user interface. Initially a coloured marker was placed on the exact location of the selected site, projected on the two dimensional coordinate system. The main features of this tool include display of:

- data availability from sites
- environmental observations
- pre-processed time series
- 3D surface visualisation

Time series visualisation and 3D gridded environmental surface visualisation was based on average daily values [10]. Three-dimensional surface visualisation was based on available patched point sensor node data. In reality some the sensor nodes not providing valid data at any given time were marked as red dot on the surface visualisation.

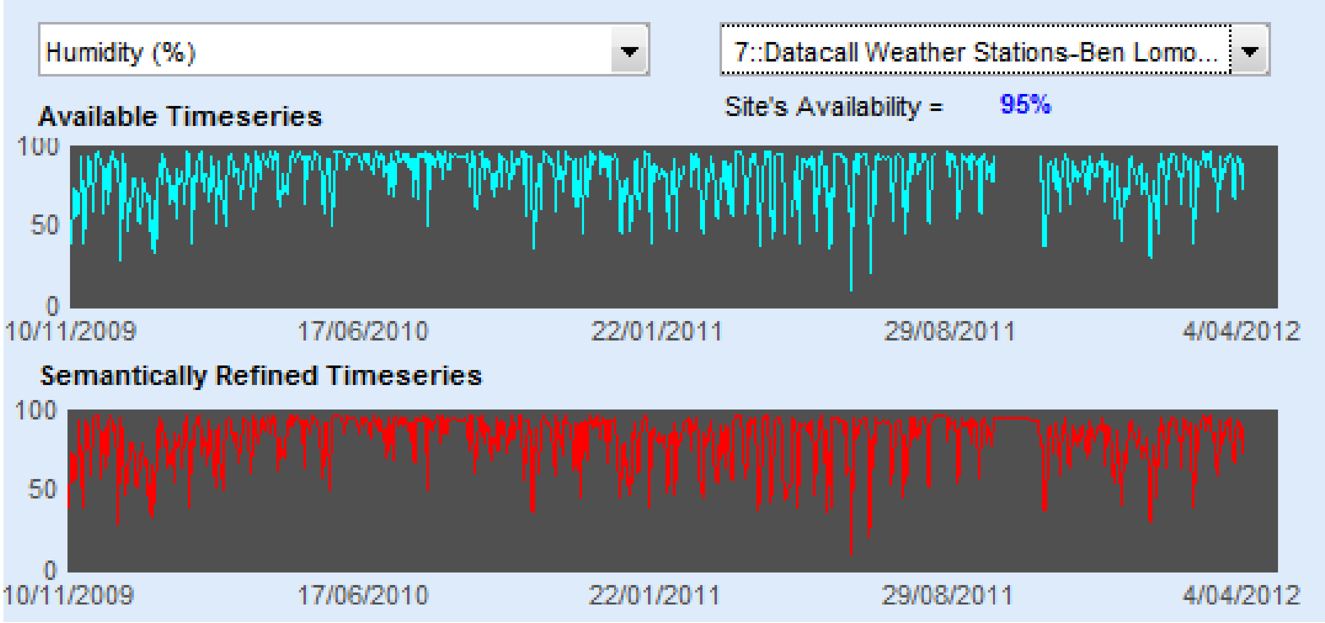


Fig. 4. Data visualisation tool developed for this study. Recorded time series and interpolated time series with missing values for complete visualisation.

B. Node Data Filtering

Pre-processing the downloaded time series was an important feature due to the uncertainty associated with data availability. Individual time series were identified according to the name of the selected site and environmental phenomena. The full time series were available since the beginning of deployment. Missing values are an unfortunate reality of nodes operating in remote, harsh operational environments and were present in these time series. For some sensor nodes, there were a number of ±Infinite values. Initially a filter was designed to remove all of the ±Infinite values, and replace them with a 'Not a Number' string to avoid introducing error and maintain the full time series length [11].

In the next stage of data pre-processing, context based filtering was applied. The valid operational ranges provided by the sensor manufacturers were used to design individual parametric filters. A sensor measuring a particular environmental parameter should operate within a well-defined range. Hence, any value recorded outside of the operational range was treated as invalid data and replaced with a 'Not a Number' string. Filtered data was stored in a structured array.

C. Data Availability Visualisation

A metric of data availability was computed as the ratio of the total number of days since a particular sensor was deployed and total number of days since a valid data point was produced. Data availability varied between 0% and 100%. Figure 3 shows the distribution of data availability for the South Esk sensor web while representing all nodes historically provided humidity data. Darker node means more historical data availability from that node.

A threshold of at least 70% or more of available data was applied. For the time series above the threshold, nearest neighbour interpolation filled in missing values. If the data availability was less than 70%, interpolation was not applied and the time series was presented with gaps. Future analysis is still possible with incomplete time series, but imputed time series are advantageous for the visualisation. This visualisation assisted in the comparison of original time series and semantic feature based refined and interpolated time series in a statistically valid way. Figure 4 contains an example comparison between raw and processed time series data recorded from a single node.

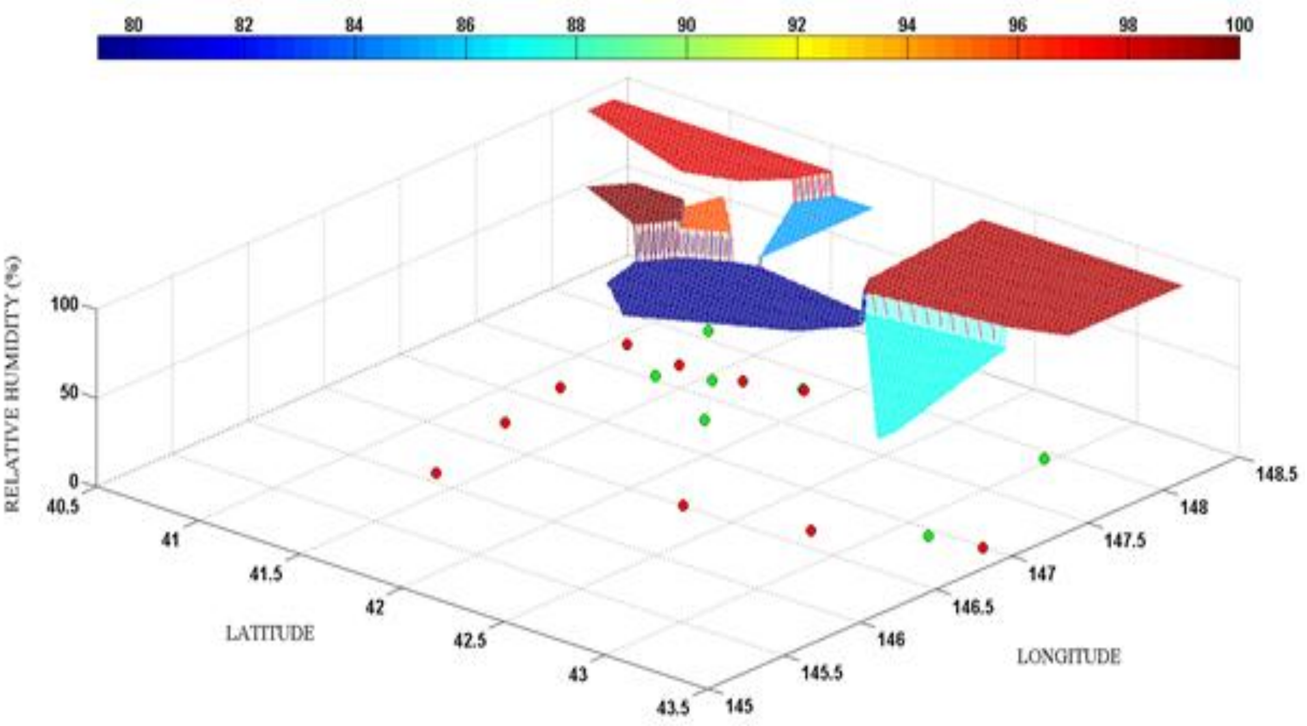


Fig. 5. 3D Gridded Mesh Relative Humidity Surface Visualisation on 19/12/2011.

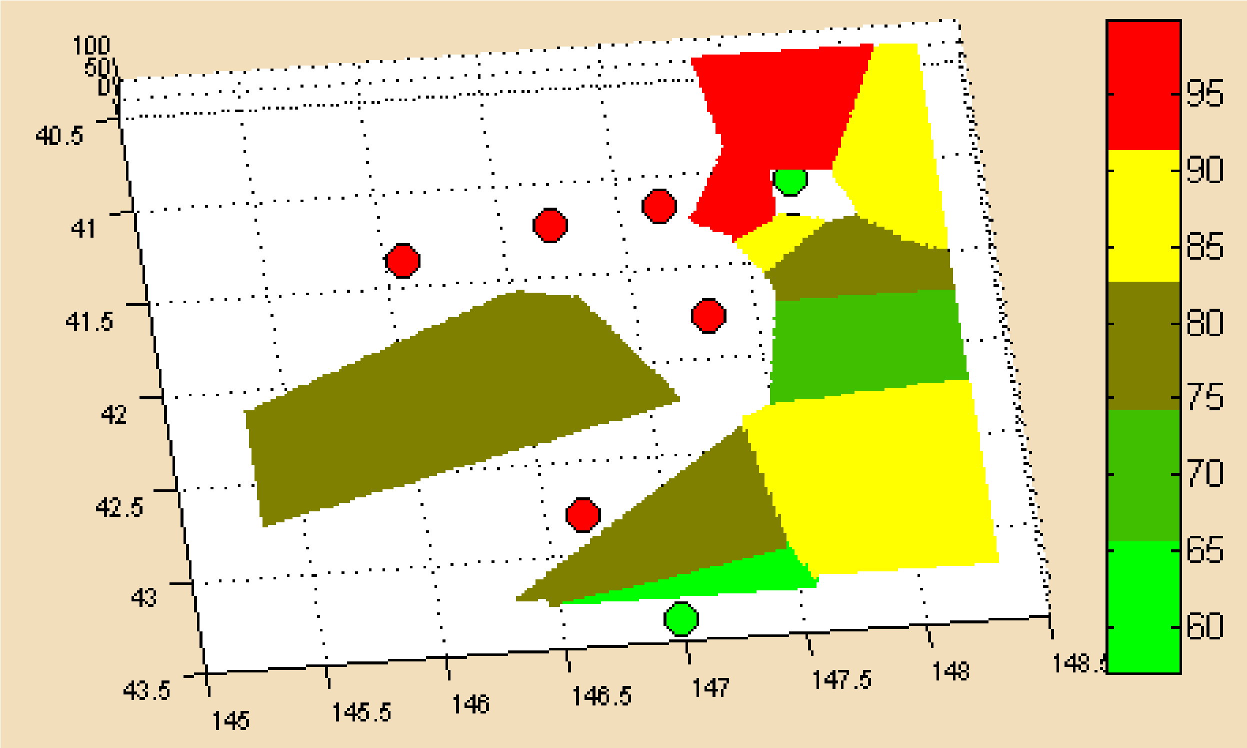


Fig. 6. 2D Gridded Mesh Relative Humidity Surface Visualisation on 19/12/2011.

D. 3D Mesh Surface Visualisation

The 3D surface visualisation was developed to provide an environmental gridded surface from the South Esk sensor web data alone. Dynamic data from 40 sensor nodes was combined to create a 3D weather surface from cubic interpolation. The natural cubic spline is a form of interpolation that uses a piecewise polynomial interplant called a spline. The benefit of spline interpolation over polynomial interpolation is that the interpolation error can be made small even when using low degree polynomials for the spline. Spline interpolation avoids the problem of Runge's phenomenon which occurs when interpolating between equidistant points with high degree polynomials [12].

For each of the environmental phenomena an individual surface was created with a daily surface generated from daily averages. Figure 5 shows the interpolated 3D gridded mesh relative humidity (%) surface for the entire South Esk sensor web. Figure 6 shows the two-dimensional (2D) view of the same visualisation. The round markers in red indicate the unavailable sensor nodes for that day and green dots markers represent nodes that provided valid data. The final surface visualisation was created using only available sensor nodes that provided valid data.

E. Dynamic Annotation and Recommendation

On the basis of data pre-processing, availability and interpolation results, a dynamic time series annotation system was developed to provide recommendations about the South Esk sensor web data. Individual time series were labelled as data quality labels, namely {'Excellent Node', 'Good Node', 'Average Node', 'Poor Node', 'Damaged Node'}. Processed time series were stored in a data structure along with recommendations.

Additional statistical features were included in the processed data, including the:

- maximum value of an event and its date
- minimum value of an event and its date
- longest missing value segment with corresponding dates
- maximum number of consecutive days with the least data variance.

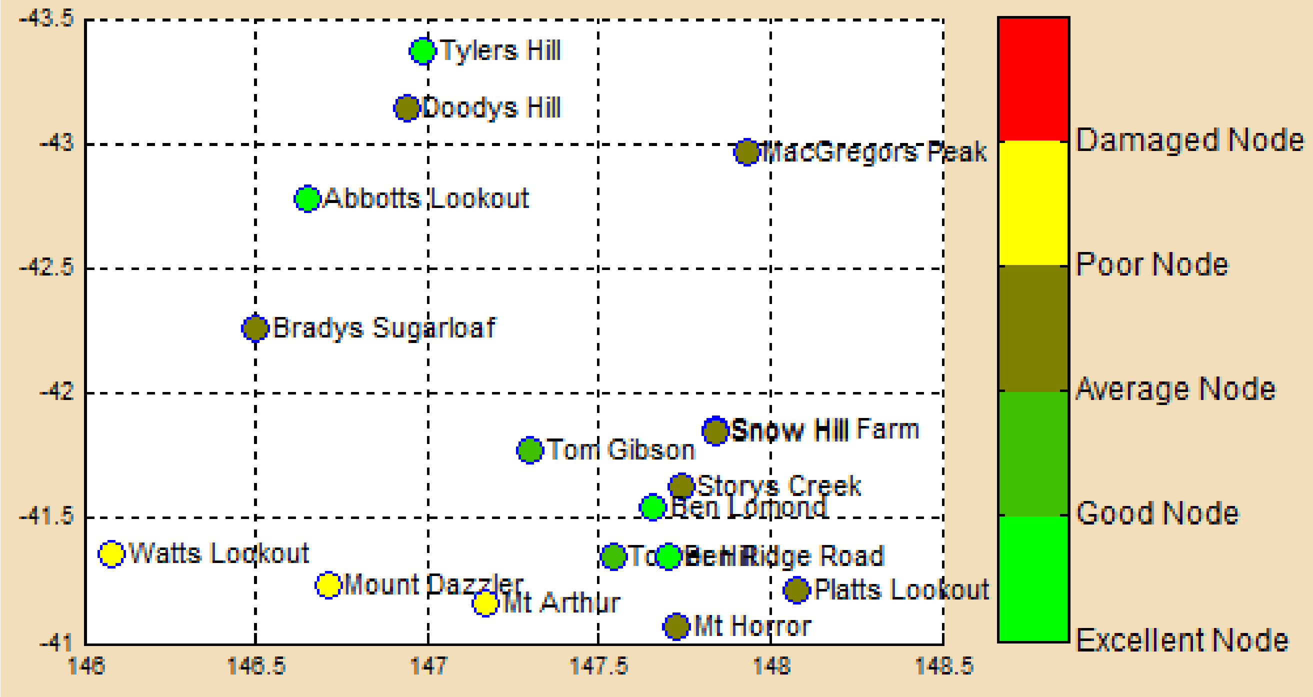


Fig. 7. Dynamic Annotation and Recommendation about the sensor network's node based data quality.

All of this processed information becomes part of the dynamic data annotation system. The purpose of this system is to process time series dynamically, annotate, and then provide a general data usability recommendation for users of the network. The recommendation of the statistical data annotation system can then assist researchers to optimize the usage of data and significantly increase the overall performance of any designed application.

Individual time series (representing individual sensor node) was labeled as one of the categories, namely, "Very Good Node (>=90%)", "Good Node (<90% and >=80%)", "Average Node (<80% and >=65%)", "Poor Node (<65% and >=50%)" and "Damaged Node (<50%)" depending on the values of the performance scores. These thresholds were defined according to the sensor specification and practical data quality experiences. Figure 7 shows the dynamic annotation and recommendation about the sensor network's node based data quality.

This data visualisation based recommendation provides a unique service, which also identifies serious issues around sensor network data quality and data delivery. The South Esk hydrological sensor web is a harsh operating environment involving very difficult terrain (including a greater than 1500 metre mountain peak), which adversely affects data acquisition and delivery from many areas of the network. Any hydrological application based upon data from this is near impossible without evaluations of reliability and data quality.

V. Conclusion

Sensor webs are macro-instruments for sensing that can be used to integrate knowledge for enhanced situational awareness in decision support applications. The integration of data from large-scale sensor webs into decision support systems requires assurances that the complementary data sources are fit for their intended purpose. We have developed the South Esk Data Service Tool to analyse and visualise the uncertainty of sensor nodes in the South Esk sensor web.

We have presented a recommendation system that provides real-time analysis of data quality and sensor reliability that can be visualised and annotated. The sensor web data can be interpolated to produce a gridded three dimensional surface of climatic variability across the South Esk. In future work we are developing new types of analysis for assessing the uncertainty of data using historical distributions. We are continuing to investigate alternative data imputation methods using the spatial context provided by correlated sensor web assets and machine learning methods.

Although this study uses the South Esk sensor web as a use case the statistical data annotation, data recommendation and sensor web visualisation could be adapted for any sensor network in the world.

Acknowledgment

The authors wish to thank the Intelligent Sensing and Systems Laboratory and the Tasmanian node of the Australian Centre for Broadband Innovation. A grant from the Tasmanian Government which is administered by the Tasmanian Department of Economic Development, Tourism and the Arts helped to conduct this research. Authors would like to acknowledge Andrew Terhorst for providing access to the South Esk sensor web. Data was delivered through the CSIRO data service system developed by Pete Taylor, Chris Peters and Brad Lee.

References

[1] A. Nasipuri, and K. Subramanian, “Development of a Wireless Sensor Network for Monitoring a Bioreactor Landfill”, GeoCongress 2006. http://www.ece.uncc.edu/~anasipur/ pubs/geo06.pdf

[2] M. Botts, G. Percivall, C. Reed, and J. Davidson. "OGC Sensor Web Enablement: Overview and High Level Architecture". GSN 2006, Lecture Notes in Computer Science, Volume 4540. pp. 175–190, 2008.

[3] M. Bartsch, T. Weiland, and M. Witting, "Generation of 3D isosurfaces by means of the marching cube algorithm”, Magnetics, IEEE Transactions on Volume 32, Issue 3, Part 1, pp.1469 – 1472, May 1996

[4] D. Smith, G. Timms, P. de Souza, and C. D’Este, “A Bayesian Framework for the Automated Online Assessment of Sensor Data Quality”, Sensors, Volume 12, Issue 7, pp. 9476-9501, 2012

[5] H.M.F. AlboElFotoh, E.S. ElMallah, and H.S. Hassanein, “On the reliability of wireless sensor networks”, Proceedings of the IEEE International Conference on Communications, pp. 3455-3460, 2006

[6] H.L. Feng and S.Y. Liu, “Reliability analysis of a wireless sensor network based on a physical model”, Journal of the Chinese Insitute of Industrial Engineers, pp. 22-27, 2010

[7] N. Purohit, P. Varadwaj, and S. Tokekar, "Reliability analysis of wireless sensor network," 16th IEEE International Conference on Networks, ICON 2008, pp.1-6, Dec 2008.

[8] The South Esk website. [Online] (2011). Available: http://www.csiro.au/sensorweb/au.csiro.OgcThinClient/OgcThinClient.html

[9] The Time series Catalog website. [Online] (2010). Available: http://www.ieee.org/http://www.csiro.au/sensorweb2/search .

[10] P. Buonadonna, D. Gay, J. M. Hellerstein, W. Hong, and S. Madden, “TASK: sensor network in a box”, Proceedings of the Second European Workshop on Wireless Sensor Networks, pp 133 -144, 2005.

[11] H. Yuxi, L. Deshi, H. Xueqin, S. Tao, H. Yanyan, “The Implementation of Wireless Sensor Network Visualization Platform Based on Wetland Monitoring”, Second International Conference on Intelligent Networks and Intelligent Systems, pp 224- 227, Nov 2009.

[12] B. Horling, R. Vincent, R. Mailler, J. Shen, R. Becker, K. Rawlins, V. Lesser, “Distributed sensor network for real time tracking”, Second International Conference on Intelligent Networks and Intelligent Systems, pp224- 227, Nov 2009.

[13] R. Dutta, D. Smith, G. Timms, “Dynamic Annotation and Visualisation of the South Esk Hydrological Sensor Web”, pp 105-110, IEEE ISSNIP, Melbourne, Australia, April 2013.

[14] R Dutta, A Morshed, “Performance Evaluation of South Esk Hydrological Sensor Web: Using Machine Learning and Semantic Linked Data Approach, Accepted in Special Issue of IEEE Sensor Journal on IoT, May 2013.

Virtual Calibration of Cosmic Ray Sensor: Using Supervised Ensemble Machine Learning

Ritaban Dutta
Intelligent Sensing and Systems Laboratory
CCI, CSIRO Hobart, Australia

Claire D'Este
Intelligent Sensing and Systems Laboratory
CCI, CSIRO Hobart, Australia

***Abstract*— In this paper an ensemble of supervised machine learning methods has been investigated to virtually and dynamically calibrate the cosmic ray sensors measuring area wise bulk soil moisture. Main focus of this study was to find an alternative to the currently available field calibration method; based on expensive and time consuming soil sample collection methodology. Data from the Australian Water Availability Project (AWAP) database was used as independent soil moisture ground truth and results were compared against the conventionally estimated soil moisture using a Hydroinnova CRS-1000 cosmic ray probe deployed in Tullochgorum, Australia. Prediction performance of a complementary ensemble of four supervised estimators, namely Sugano type Adaptive Neuro-Fuzzy Inference System (S-ANFIS), Cascade Forward Neural Network (CFNN), Elman Neural Network (ENN) and Learning Vector Quantization Neural Network (LVQN) was evaluated using training and testing paradigms. An AWAP trained ensemble of four estimators was able to predict bulk soil moisture directly from cosmic ray neutron counts with 94.4% as best accuracy. The ensemble approach outperformed the individual performances from these networks. This result proved that an ensemble machine learning based paradigm could be a valuable alternative data driven calibration method for cosmic ray sensors against the current expensive and hydrological assumption based field calibration method.**

Keywords—Cosmic Ray sensor; Ensemble supervised machine learning; Area wise bulk soil moisture.

I. COSMOZ DATA AND SYSTEM

The Australian Cosmic Ray Sensor Soil Moisture Monitoring Network (CosmOz) (Figure 1) [1-2] is a near-real time continental scale soil moisture monitoring system originally inspired by the United States Cosmic-ray Soil Moisture Observing System (COSMOS) [3-5]. CosmOz aims to test the utility of Hydroinnova CRS-1000 cosmic ray soil moisture probes [3] (Figure 2) for water management, water information, hydrological process research applications and test the feasibility and utility of a national near-real time soil moisture measurement network.

The cosmic ray soil moisture probe measures the neutrons released when cosmic rays interact with hydrogen atoms in water molecules found in the soil and atmosphere. The number of fast neutrons emitted into the atmosphere is inversely correlated with soil moisture. Figure 3 shows the fundamental principal behind these cosmic ray probes [5-6].

Data from the Hydroinnova CRS-1000 cosmic ray soil moisture probe deployed in the Tullochgorum site in Tasmania, Australia was used for this study. It consists of two neutron detectors: a bare detector that responds mainly to thermal neutrons and a polyethylene-shielded detector that responds mainly to epithermal-fast neutrons. Each counter has its own high-voltage power supply and a pulse module to analyse the signal generated by the neutron detector tube. An Iridium satellite modem then transmits the data at one hour time intervals to the CosmOz CSIRO data server [7-8].

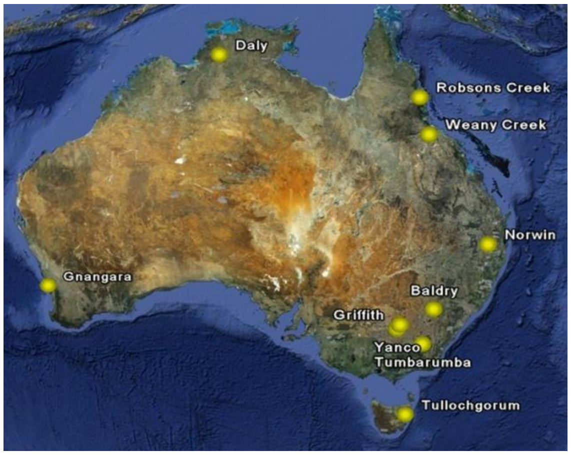


Fig. 1. CosmOz has already deployed Hydroinnova CRS-1000 cosmic ray soil moisture probes at 11 different locations throughout Australia.

Tullochgorum CosmOz cosmic ray sensors were calibrated using soil samples collected around the probe. Soil moisture was measured using the oven-drying method, and area-average soil moisture was computed for all samples [9]. Finally the average soil moisture content within the footprint of the probe was used to convert neutron counts into soil moisture [10-11]. Figure 4 shows the rainfall, pressure corrected neutron count profile, and conventionally estimated soil moisture profile for the period chosen for this study (July 2011 to May 2013) at the Tullochgorum site.

The Intelligent Sensing and Systems Laboratory, CSIRO CCI and the Tasmanian node of the Australian Centre for Broadband Innovation are assisted by a grant from the Tasmanian Government which is administered by the Tasmanian Department of Economic Development, Tourism and the Arts.

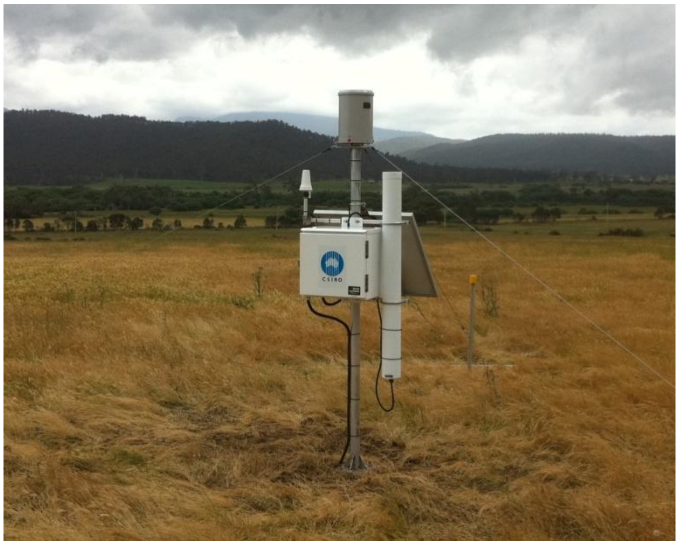


Fig. 2. The Australian Cosmic Ray Sensor Network‘s Hydroinnova CRS-1000 cosmic ray soil moisture probe deployed at the Tullochgorum site.

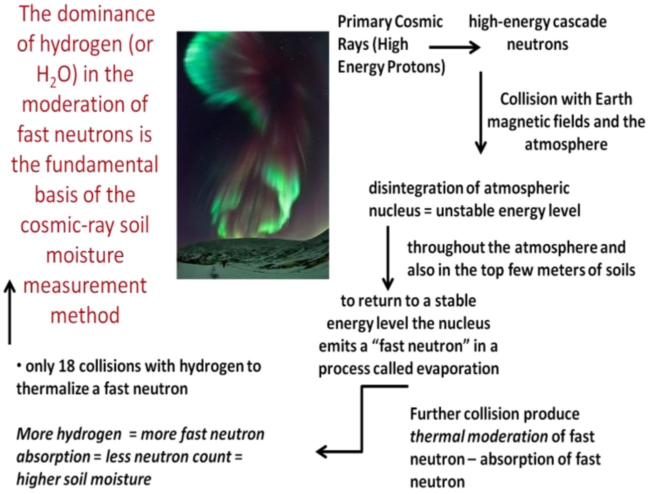

The dominance of hydrogen (or H_2O) in the moderation of fast neutrons is the fundamental basis of the cosmic-ray soil moisture measurement method
Primary Cosmic Rays (High Energy Protons)
high-energy cascade neutrons
Collision with Earth magnetic fields and the atmosphere
disintegration of atmospheric nucleus = unstable energy level
throughout the atmosphere and also in the top few meters of soils
to return to a stable energy level the nucleus emits a "fast neutron" in a process called evaporation
• only 18 collisions with hydrogen to thermalize a fast neutron
Further collision produce thermal moderation of fast neutron – absorption of fast neutron
More hydrogen = more fast neutron absorption = less neutron count = higher soil moisture

Fig. 3. Schematic flow diagram of the fundamental principal behind this cosmic ray probes.

II. Calibration Problem Space

The current field calibration method for cosmic ray sensors is significantly time consuming and expensive. For the current method, it is also essential to conduct at least two field calibrations (dry calibration during summer and wet calibration during winter) to complete the minimum requirement for of the calibration process. Ideally monthly calibration is required, which makes the current calibration process very expensive and impractical. Within the current calibration method, corrective equations are presently used to nullify the effects of universal cosmic radiation, lattice water in soil, pressure and humidity. However, there are uncertainties and questions about the effectiveness of these calibration equations in order to produce real-time accurate soil moisture estimations. The process is prone to human error as it is based on limited experimental data, hard manual field sampling protocol, and theoretical assumptions [1-9].

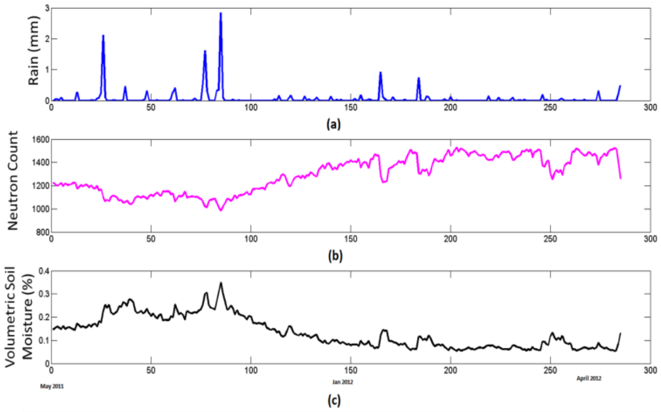


Fig. 4. CosmOz Data from the period July 2011 – May 2013 were used for testing and validation for this research study, (a) Daily rain fall, (b) Pressure corrected neutron counts, (c) Estimated bulk soil moisture profile for Tullochgorum site – converted from pressure corrected neutron count data generated by the Hydroinnova CRS-1000 cosmic ray soil moisture probe..

The main focus of this study was to investigate the possibilities of a complete data driven virtual sensor calibration approach which could be well suited for this purpose. We proposed an ensemble machine learning based alternative method to capture the behavioural aspect of the neutron count observations compared with the real soil moisture measurements. The aim was to develop a virtual calibration method, cross validated against an external ground truth data source, independent of CosmOz network. The Australian Water Availability Project (AWAP) data base [12] has been used for this purpose.

III. Experimental Data Sets Design

The AWAP soil moisture data base [12] was used as an external data source, independent of CosmOz network to develop the virtual calibration method [2]. The AWAP database monitors the state and trend of the terrestrial water balance of the Australian continent, using model-data fusion methods to combine both measurements and modeling. The AWAP database provides 16 environmental attributes from which Radiation (MJ/m2), Max Temperature (degC), Min Temperature (degC), Rainfall (mm), Soil Evaporation (mm), Local Discharge (Runoff+Drainage) (mm), Surface Runoff (mm), Open Water Evaporation ('pan' equiv) (mm), Deep Drainage (mm), Sensible Heat Flux (MJ/m2) and Latent Heat Flux (MJ/m2). These attributes were used as part of the training and testing input for the machine learning algorithms. Selection of these inputs from the whole AWAP data set was based on expert domain knowledge and principal component analysis (PCA) as feature selection method, where eleven least correlated attributes (covering 99% of data variance) were selected for the experimental design. Figure 5 shows the training and testing inputs from AWAP. The pressure corrected fast neutron count time series from CosmOz was also considered as part of the whole training and testing input sets (Figure 4). On the other hand Upper Layer Soil Moisture (%) and Lower Layer Soil Moisture (%) were used as the training target for the data experiments (Figure 5). An AWAP data adaptor was developed and used to download and unzip all

AWAP folders, process the sequential NetCDF gridded data files, and to extract all the time series. Various training and testing data sets were formed based on a randomized incremental optimization algorithm to evaluate the generalization capability of the proposed individual and ensemble machine learning architectures. Combinations of % training data and % testing data were varied from {10%-90%} to {50%-50%} to identify the best possible training-testing data balance to achieve maximum prediction accuracy with highest possible sensitivity and specificity.

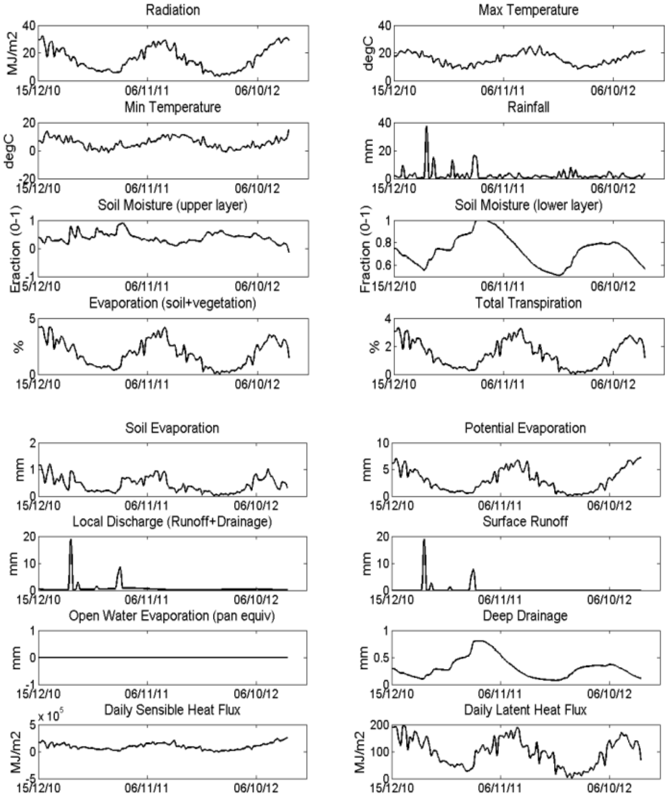


Fig. 5. Individual time series were extracted from the AWAP gridded maps based on pixel values corresponding to the latitude- longitude information of Tullochgorum site in Tasmania, Australia. Various training and testing data sets were formed based on a randomized incrementing optimization algorithm.

IV. SUPERVISED ESTIMATORS

Four supervised estimators, namely Sugano type Adaptive Neuro Fuzzy Inference System (S-ANFIS) [17-18], Cascade Forward Neural Network (CFNN) [19-21], Elman Neural Network (ENN) [22-23] and Learning Vector Quantization Neural Network (LVQN) [24-25] were selected for this study. In previous work related to this topic S-ANFIS was used for soil moisture estimations [1].

A. *Rationale*

The main rationale behind selecting these four supervised estimators was to conduct a comparative study on significantly varied neural network architectures predicting soil moisture based on the same fast neutron counts to identify a better architecture for this calibration purpose [26-27]. In the later stage of this paper, an ensemble approach has been proposed for better soil moisture estimation and dynamic cosmic ray probe calibration. In an ensemble approach selection of widely varied supervised estimators was another essential rationale to achieve much better prediction generalization and higher calibration accuracy. Ultimate goal was to have a parallel processing of the neutron count data using multiple ANNs to capture significant data behavioural variance in relation to the soil moisture and also in the predicted time series; so that ensemble generalization can perform better than one individual ANN, hence they could complement each other on a dynamic range. The MATLAB programming environment was used to train and test these estimators. Individual performances were evaluated based on the common training and testing paradigms at any given time.

B. *Performance Evaluation*

Performance assessment was conducted using a point to point comparative study between the soil moisture time series output from the trained estimators during testing and the soil moisture time series output obtained from the existing field calibration methodology. Higher percentage similarity between these two time series was essential to justify the effectiveness of the individual estimator based alternative method before further improvement could be achieved. Each point on the both time series was representing a single day, so point to point comparison provided a daily comparison.

A predictive performance estimation mechanism based on time series cross correlation and auto correlation was applied to measure percentage accuracies of the predictions. High correlation between expected soil moisture profile and the predicted one represented better prediction performances. Finally, performances of these estimators were quantified by prediction accuracy ((TP + TN) / (TP + FN + FP + TN) where true positives =TP, true negatives =TN, false positives = FP, false negatives = FN). The evaluation process also included sensitivity (TP / (TP + FN)); specificity (TN / (FP + TN) calculations to justify the estimation correctness. As in hydrology ± 3% tolerance limit is acceptable in soil moisture measurements, point to point comparison between two time series provided us TP, TN, FP, and FN estimates [4-6].

C. *S-ANFIS Estimator*

S-ANFIS is a neural network method based on the Takagi–Sugeno fuzzy inference system. Since it integrates both neural networks and fuzzy logic principles, it has the potential to capture the benefits of both in a single framework [17]. S-ANFIS uses only differentiable functions thus standard learning procedures from neural network theory can easily be used. The parameters are propagated again, and in this epoch back-propagation is used to modify the antecedent parameters or the membership functions, while the consequent parameters remain fixed [18]. The generation of the rule base is unsupervised followed by supervised learning to update the rule parameters. In this study the supervised part of the S-ANFIS estimator was a multi-layered perceptron network (MLPN). A sigmoid activation function in the form of a hyperbolic tangent has been used in this estimator.

D. *CFNN Estimator*

CFNN is similar to feed-forward networks, but include a connection from the input and every previous layer to the following layers. As with feed-forward networks, a two-or more layered cascade-network can learn any finite input-output relationship arbitrarily and predict time series sequences well; provided it has enough hidden neurons [17, 19-21]. Sigmoid activation function with normalization between -1 and 1 has been used.

E. ENN Estimator

The ENN is a simple recurrent neural network consisting of an input layer, a hidden layer, and an output layer. In this way it resembles a three layer feedforward neural network. Elman neural networks are very useful for predicting time series sequences, since they have a limited short-term memory [17, 22-23]. Short-term memory provides a unique capability to the ENN, by storing the previous learning step which could be used to influence the next learning step. At each time step, the input is propagated in a standard feed-forward fashion, and then a learning rule is applied. The fixed back connections result in the context units always maintaining a copy of the previous values of the hidden units (since they propagate over the connections before the learning rule is applied). Thus the network can maintain a sort of state, allowing it to perform such tasks as sequence-prediction that is beyond the power of a standard multilayer perceptron.

F. LVQN Estimator

A LVQNN consists of two layers competitive layer and linear layer. The first layer maps input vectors into clusters that are found by the network during training. The second layer merges groups of first layer clusters into the classes defined by the target data. The total number of first layer clusters is determined by the number of hidden neurons. The larger the hidden layer the more clusters the first layer can learn, and the more complex mapping of input to target classes can be made [17, 24-25].

V. GENERALISATION RESULTS AND DISCUSSION

The CFNN network required typically 2,000 training iterations, ENN required only 750 training iterations and LVQNN needed only 1400 training iterations. The CFNN (with learning rate equal to 0.42 and a momentum term equal to 0.5) with eleven inputs, ten hidden and one output neuron was able to reach a best success rate of 84.3% (rates varied between 67% - 84.3% for various training-testing paradigms as describe in section 3) in correct soil moisture prediction while using {75% training – 25% testing} paradigm. The ENN with same architecture (with an additional recurrent layer with tap delay 1:5, where hidden later size was 50, and 'trainlm' as training function) was able to reach a best success rate of 86% (rate varied between 75% - 86%) while using {70% training – 30% testing} paradigm. In the LVQNN, neurons were added to the network until the sum-squared error (SSE) falls beneath an error goal (0.001), or a maximum number (172) of internal neurons was reached. It was important that the spread parameter was large enough so that the hidden neurons respond to overlapping regions of the input space, but not so large that all the neurons respond in essentially the same manner. The spread parameter was set to 0.79. The LVQNN was able to achieve a maximum of 80% prediction accuracy using {65% training – 35% testing} paradigm. S-ANFIS training was fastest among all of the estimators. S-ANFIS was able to predict bulk soil moisture at an accuracy of 87% while using {80% training – 20% testing} paradigm (accuracy varied from 72% - 87%).

TABLE I. COMPARATIVE PREDICTION ACCURACY RESULTS

Optimum Experimental Paradigm where Best Individual Performances were Recorded	**S-ANFIS % Accuracy**	**CFNN % Accuracy**	**ENN % Accuracy**	**LVQNN % Accuracy**
{75% training – 25% testing}	83	***84.3***	78.4	73
{70% training – 30% testing}	76.5	67	***86***	77
{80% training – 20% testing}	***87***	75	75	74
{65% training – 35% testing}	72	79.5	76.3	***80***
Best Sensitivity Recorded (%)	81	75	73	72
Best Specificity Recorded (%)	76	84	70	79
Best False Positive (%)	6.91	12	9.23	11.2
Best True Negative (%)	89.5	87.6	85	82

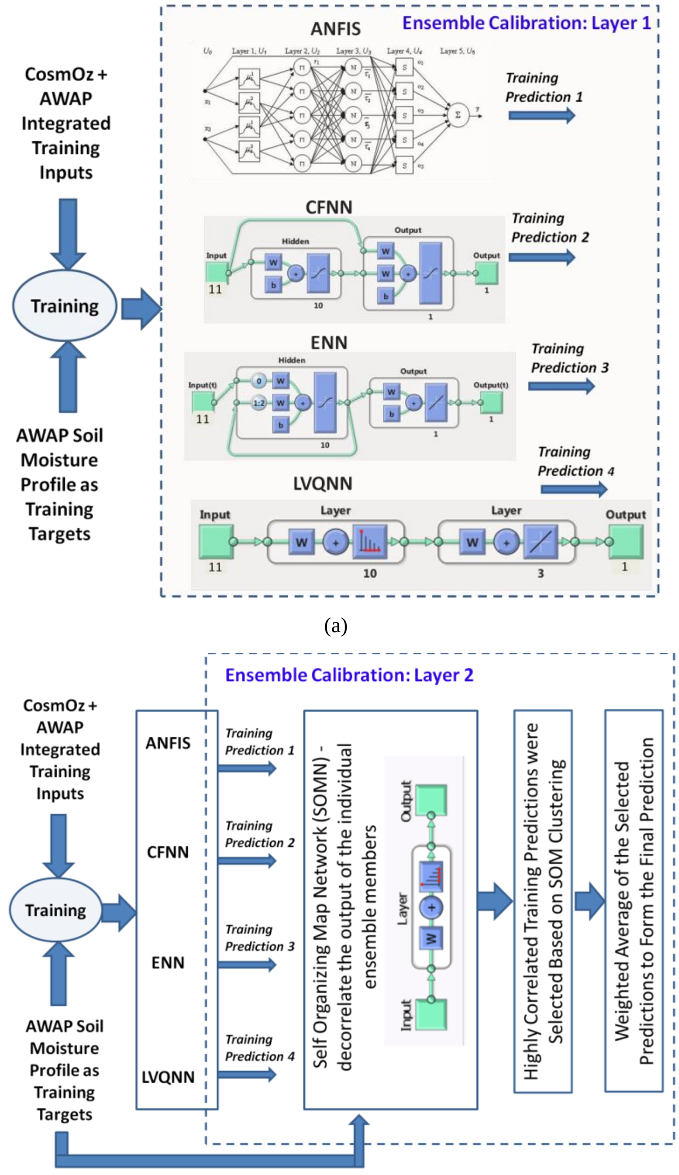
S-ANFIS was the overall best performer compared to the other three estimators whereas LVQNN was able achieves maximum prediction accuracy with least amount of data being used from training. TABLE 1 summarizes all the generalization results using all four SML estimators in terms of correct percentages of soil moisture prediction. Results are presented for four different experimental scenarios with different combination of training-testing, where individual supervised estimator had maximum prediction accuracy. It was evident that although prediction accuracies were significantly high, but there was no best architecture as individual estimator’s performance sensitivity and specificity were quite mixed with no clear winner. Ultimately supervised neural network based alternative virtual calibration could only be useful if a better prediction accuracy along with higher sensitivity and specificity could be achieved than what is recorded in the TABLE 1.

At this point it was a natural progression to apply an ensemble learning paradigm where several supervised estimators could be jointly used to calibrate cosmic ray sensors. In the next section two layered ensemble approach was applied to combine and complement these four estimators in order to explore any possible improvement in soil moisture estimation and virtual calibration of these cosmic ray sensors.

VI. Supervised Ensemble Approach

A. Two Layered Ensemble Methodology

Two layered ensemble approach has been proposed and developed in this study to explore the possibility of overall improvement. In the first layer global training was performed on all four supervised estimators (S-ANFIS, CFNN, ENN and LVQNN) using a common training and target set. Once trained all the estimators were simulated (tested) using the same training inputs, but without the targets to generate four individual predictions.


Ensemble Calibration: Layer 1
CosmOz + AWAP Integrated Training Inputs
Training
AWAP Soil Moisture Profile as Training Targets
ANFIS
Training Prediction 1
CFNN
Hidden
Output
Input
11
10
1
Training Prediction 2
ENN
Input(t)
Output(t)
Training Prediction 3
LVQNN
Layer
3
Training Prediction 4
Ensemble Calibration: Layer 2
ANFIS
CFNN
ENN
LVQNN
Training Prediction 1
Training Prediction 2
Training Prediction 3
Training Prediction 4
Self Organizing Map Network (SOMN) - decorrelate the output of the individual ensemble members
Highly Correlated Training Predictions were Selected Based on SOM Clustering
Weighted Average of the Selected Predictions to Form the Final Prediction

(a)

(b)

Fig. 6. (a) The schematic design for the layer 1 of the proposed ensemble supervised machine learning approach for cosmic ray sensor calibration, where individual supervised estimators are being trained in parallel; (b) Second layer of the ensemble approach, where SOMN is applied to decorrelate the training predictions against the training targets to select the highly correlated ones, and throw away the least correlated predictions.

The next layer of this proposed approach was constructed using a self-organizing map network (SOMN) and weighted averaging block. A self-organizing map (SOMN) is a type of artificial neural network that is trained using unsupervised learning to produce a low-dimensional (typically two-dimensional), discretized representation of the input space of the training samples, and called a map. SOMN may be considered a nonlinear generalization of PCA, where cross correlation among various inputs get projected on an empirical orthogonal plane based proportion of data variance captured along different components. SOMN was selected to decorrelate the predictions from the ANNs as it has more generalization capability to cover data variance than PCA or any cross correlation based methods, so less amount of data variance remains unexplained [33-34]. Individual ANN is providing single dimensional prediction. Four predicted time series from four different estimators are combined as inputs to the SOM.

Once trained initial SOM grid (the 100 neurons with network size 10 X 10) was re-distributed among the data points, distributed concentration of the trained neurons formed natural separated clusters. Two clusters were highly correlated if they were comparatively closely positioned - hence they were selected. Least correlated ones were thrown out. An objective function based on intra cluster distance measure was used to determine the cluster correlations (Figure 7).

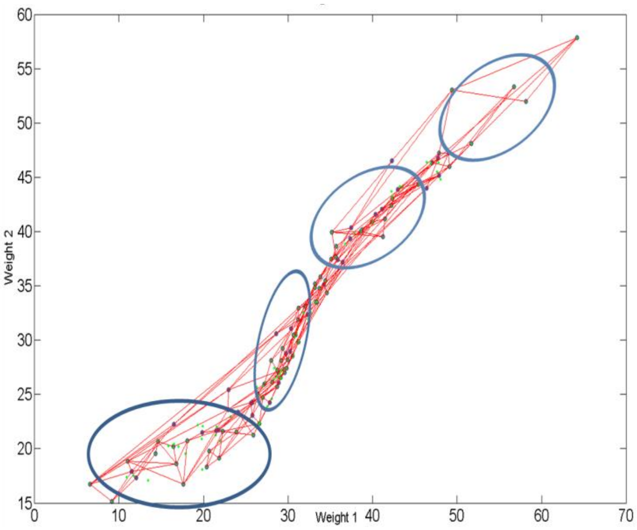


Fig. 7. The representation of initial cluster positioning of the predicted time series from the four different supervised estimators. SOMN was applied to decorrelate these clusters to select fewer highly correlated clusters (typically first two in most of the cases).

SOM decorrelate these inputs to create individual pattern maps (and clusters) for individual input time series based on training weights (Fig. 8). Based on the SOMN natural clustering (or natural grouping of the predicted values) on the predictions from the layer 1 and the training targets, highly correlated ones were selected as they had statistically similar fluctuations. Euclidian distance among the SOM clusters were used to define the number of similar clusters hence highly correlated. Data points which belonged to the similarly positioned clusters were then marked and traced back to flag the corresponding original supervised estimators together as highly correlated. Data points belong to those selected clusters were then traced back to the corresponding ANN methods, and predictions from those methods were selected for averaging to form the final one dimensional prediction. Initial prediction values proposition in an initialized SOMN are depicted in Figure 8.

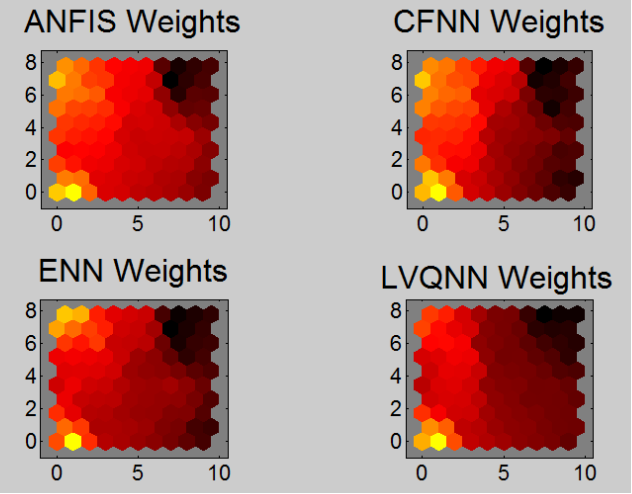


Fig. 8. Example of trained SOMN based de-correlated individual weight maps for the four supervised estimators. In this instance predictions from ANFIS and LVQNN have been selected for final stage of prediction estimation, while 80% data is used for training.

The average of the output of several different models $f_i(x)$could be called as an ensemble model which will take the form of $\widehat{f(x)} = \sum_{i=1}^{K} w_i f_i(x)$. The idea of averaging different outputs from different models was developed in the neural network community [28-29]. Later it was also established by Krogh et al. [30] that the generalization error of the ensemble could be lower than the mean of the generalization error of the single ensemble members. It has also been reported in the literature that the generalization error of an ensemble model could be improved if the predictors on which averaging is done disagree and if their fluctuations are uncorrelated [31-32]. Development of the ensemble approach was primarily focused on establishing a mechanism where decomposition of several predictions could be done in order to establish the highly correlated ones and throw away the uncorrelated ones.

Selected predictions were then averaged to form the final prediction. Sorted corresponding cross correlation coefficients were used to perform the weighted average. Based on the SOM weight maps correlation coefficient between [-1 and +1] were generated. Ensembles with higher coefficients are given higher weighting in the prediction. This additional processing was used to refine the final outcome in more realistic way. First two highly correlated predictions were used to perform the weighted average to form the final outcome. Based on the nature of dynamically available time series different estimators were selected based on SOM based de-correlation and selection processes.

B. *Ensemble Performance Evaluation*

Performance of this newly proposed ensemble architecture was evaluated using rigorous testing paradigms. As described in section 3 amount of data used in testing was varied from 10% to 50%. Evaluation of the generalization performance concluded that the best bulk soil moisture prediction was achieved while 30% data was used for testing with 70% being used for training. Selected ANNs during ensemble training phase were then used in the testing phase as it was obvious that they had the best generalization capabilities for that particular training-testing paradigm instance. Accuracies were calculated based on the final testing prediction and the ground truth testing targets. Overall accuracy was 94.4% with 91% sensitivity, 90% specificity, 2% false positive and 95% true negative. Results based on ensemble show a clear improvement from the prediction point of view, which also proved the intended effectiveness of the proposed ensemble approach. Figure 9 shows the prediction performance based on ensemble approach.

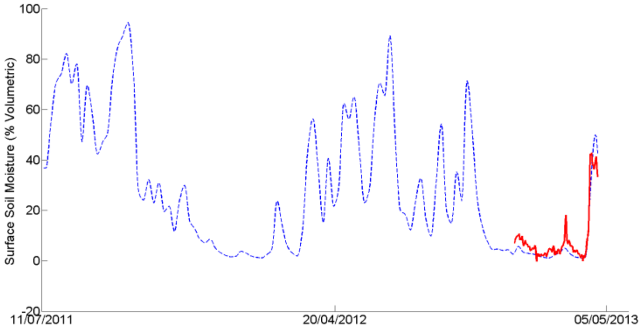


Fig. 9. Ensemble approach based soil moisture prediction performance, (red curves while blue represents testing target ground truth).

C. *Ensemble Calibration Evaluation*

Based on ensemble performance evaluation results it was evident that ensemble approach could be able to provide a unique platform for calibrating cosmic ray sensors on a dynamic basis. Two layered ensemble approach could be used as frequently as possible to train and capture new neutron counts data to estimate soil moisture profile. Higher performance accuracy with very low false positive results during testing phase shows the high level of reliability on this approach. Ensemble based approach is a virtual approach with high degree of flexibility which offers high frequency remote sensor calibration compared to the expensive field soil sampling method. This dynamic machine learning based virtual calibration could be a benchmarking methodology for calibrating cosmic ray sensors measuring area wise bulk soil moisture.

VII. CONCLUSION

This study concluded the ensemble of supervised machine learning algorithms could be an effective alternative calibration method for remote area wise estimation of bulk soil moisture using the cosmic ray sensor’s fast neutron count readings. Using the AWAP database it was possible to train the ensemble supervised estimator with historical ground truth soil moisture data, which provided better generalization capability to predict accurate soil moisture from the cosmic neutron counts. Prediction results were very encouraging. Potentially this could help us to develop a web based remote virtual sensor calibration mechanism. This way the cosmic ray sensor could be monitored and calibrated virtually and continuously.

ACKNOWLEDGMENT

The authors wish to thank the Intelligent Sensing and Systems Laboratory and the Tasmanian node of the Australian Centre for Broadband Innovation. Authors would like to thanks Andrew Terhorst, Peter Briggs, Rex Keen, Bill Cotching, Dave McJannet, Aaron Hawdon and Peter Taylor for their support.

REFERENCES

[1] R. Dutta, A. Terhorst, A. Hawdon, B. Cotching, Bulk Soil Moisture Estimation Using CosmOz Cosmic Ray Sensor and ANFIS, IEEE Sensors 2012 Proceedings, Taipei, Taiwan, 978-1-4577-1767-3/12, (2012) , 741 -744.

[2] R Dutta, A Terhorst, "Adaptive Neuro-Fuzzy Inference System Based Remote Bulk Soil Moisture Estimation: Using CosmOz Cosmic Ray Sensor", IEEE Sensors Journal, Volume 13 , Issue 6, pp. 2374 – 2381, 2013.

[3] http://hydroinnova.com/main.html Accessed July 2013.

[4] T.E. Franz, M. Zreda, R. Rosolem, T.P.A Ferre, A universal calibration function for determination of soil moisture with cosmic-ray neutrons. Hydrology and Earth System Sciences 17, (2013), 453-460.

[5] B. Hornbuckle, S. Irvin, T. E. Franz, R. Rosolem, and C. Zweck, The potential of the COSMOS network to be a source of new soil moisture information for SMOS and SMAP, paper presented at Proc. IEEE Intl. Geosci. Remote Sens. Symp., Munich, Germany, 2012.

[6] T. E. Franz, M. Zreda, T.P.A Ferre, R Rosolem, C. Zweck, S. Stillman, X. Zeng, W. J. Shuttleworth, Measurement depth of the cosmic ray soil moisture probe affected by hydrogen from various sources. Water Resources Research 48, (2012).

[7] C.A. Rivera Villarreyes, G. Baroni, S. E. Oswald, Integral quantification of seasonal soil moisture changes in farmland by cosmic-ray neutrons. Hydrology and Earth System Sciences 15, (2011), 3843-3859.

[8] P. Carlson, A century of cosmic rays. Physics Today, 65, (2012), 30–36.

[9] D. Desilets, M. Zreda, Nature's neutron probe: land-surface hydrology at an elusive scale with cosmic rays. Water Resour. Res., 46, (2010), W11505, DOI: 10.1029 /2009WR008726.

[10] D. Desilets, M. Zreda, Spatial and temporal distribution of secondary cosmic-ray nucleon intensities and applications to in-situ cosmogenic dating. Earth Planet. Sc. Lett., 206, (2003), 21–42.

[11] E. Fermi, Artifical radioactivity produced by neutron bombardment, Nobel Prize lecture, (1938), 414–421.

[12] http://www.eoc.csiro.au/awap/ AWAP Website Accessed March 2013.

[13] R. R. Brooks, S. S. Iyengar, Multi-Sensor Fusion: Fundamentals and Applications with Software. Prentice Hall PTR, Upper Saddle River, New Jersey 07458, ISBN 0-13-901653-8

[14] S. B. Kotsiantis, I. D. Zaharakis, and P. E. Pintelas. Supervised machine learning: A review of classification techniques, Frontiers in Artificial Intelligence and Applications 160 (2007): 3.

[15] T. Dietterich, Ensemble methods in machine learning, Multiple classifier systems (2000): 1-15.

[16] T. G. Dietterich, Approximate statistical tests for comparing supervised classification learning algorithms, Neural computation 10.7 (1998): 1895-1923.

[17] www.mathworks.com Accessed July 2013.

[18] J.-S.R. Jang, ANFIS: adaptive-network-based fuzzy inference system. IEEE Transactions on Systems, Man and Cybernetics, Volume: 23 , Issue: 3, (1993), 665 – 685.

[19] P. P. V. D Smagt. "Minimisation methods for training feed-forward networks." Neural Networks 7.1 (1994): 1-11.

[20] S. Singhal, and W. Lance, Training feed-forward networks with the extended Kalman algorithm, IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP-89, (1989).

[21] J. J. Hopfield, Learning algorithms and probability distributions in feed-forward and feed-back networks, Proceedings of the National Academy of Sciences 84.23, (1987), 8429-8433.

[22] J. L. Elman, Learning and development in neural networks: The importance of starting small, Cognition 48.1 (1993): 71-99.

[23] X. Z. Gao, X. M. Gao, and S. J. Ovaska, A modified Elman neural network model with application to dynamical systems identification, IEEE International Conference on Systems, Man, and Cybernetics, Vol. 2. (1996).

[24] S. C. Ahalt et al, Competitive learning algorithms for vector quantization, Neural networks 3.3 (1990): 277-290.

[25] P. Schneider, M. Biehl, and B. Hammer, Adaptive relevance matrices in learning vector quantization, Neural Computation 21.12 (2009): 3532-3561.

[26] R. Dutta, J.W. Gardner, E.L. Hines, "Classification of ear, nose, and throat bacteria using a neural-network-based electronic nose", MRS bulletin 29 (10), 709-713.

[27] A. Das, N.G. Stocks, A. Nikitin, E.L. Hines, "Quantifying stochastic resonance in a single threshold detector for random aperiodic signals", Fluctuation and Noise Letters 4 (02), L247-L265, 2004.

[28] L. Hansen, P. Salamon, "Neural Network Ensembles," IEEE Trans.in Pattern Analysis and Machine Intelligence, vol. 12, no. 10, pp. 993–1001, 1990.

[29] M. P. Perrone, L. N. Cooper, "When Networks Disagree: Ensemble Methods for Hybrid Neural Networks," in Neural Networks for Speech and Image Processing, R. J. Mammone, Ed. Chapman-Hall, 1993, pp.126–142.

[30] [30] A. Krogh, J. Vedelsby, "Neural network ensembles, cross validation, and active learning," in Advances in Neural Information Processing Systems, G. Tesauro, D. Touretzky, and T. Leen, Eds., vol. 7. The MIT Press, 1995, pp. 231–238.

[31] [31] A. Krogh, P. Sollich, "Statistical mechanics of ensemble learning," Physical Review E, vol. 55, no. 1, pp. 811–825, 1997.

[32] [32] J¨org D. Wichard, Maciej Ogorzałek, "Time Series Prediction with Ensemble Model," Neural Networks, 2004. Proceedings. 2004 IEEE International Joint Conference on Neural Networks, vol.2, pp.1625 - 1630.

[33] [33] Y Liu, R. H. Weisberg, "Patterns of Ocean Current Variability on the West Florida Shelf Using the Self-Organizing Map", Journal of Geophysical Research, 110, 2005, doi:10.1029/2004JC002786.

[34] [34] Y. Liu, R. H. Weisberg, C. Mooers, "Performance Evaluation of the Self-Organizing Map for Feature Extraction", Journal of Geophysical Research, 111, 2006 doi:10.1029/2005jc003117.

Debranding in Fantasy Realms: Perceived Marketing Opportunities within the Virtual World

Kear Andrew, Bown Gerald Robin, Christidi Sofia.
Faculty of Business, Management, Accounting and Law.
University of Gloucestershire, Gloucestershire, UK

Abstract— **This paper discusses the application of the concept of debranding within immersive virtual environments. In particular the issue of the media richness and vividness of experience is considered in these experience realms that may not be conducive to traditional branding invasive strategies. Brand equity is generally seen to be the desired outcome of branding strategies and the authors suggest that unless the virtual domains are considered as sacred spaces then brand equity may be compromised. The application of the above concepts is applied to the differing social spaces that operate within the different experience realms. The ideas of resonance, presence and interactivity are considered here. They lead to the development of a constructed positioning by the participants. Through the process of debranding, marketers may be able to enter these sacred spaces without negative impact to the brand. Perception of these virtual spaces was found to be partially congruent with this approach to branding. It thus presents a number of challenges for the owners of such virtual spaces and also virtual worlds in increasing the commercial utilization of investment in these environments.**

Keywords-component; Fantasy Realms, Debranding; Sacred Space;Marketing; 3D Virtual Worlds; Second Life

I. Introduction (Marketing Opportunity in Fantasy Realms)

The core focus of this paper is to assess the perceptions of opportunities within immersive virtual environments (IVEs). The particular environments discussed are the Fantasy realms within Second Life. The entrants to these fantasy realms develop a community of participation. Marketing opportunity is contingent upon the perception and selection of the optimum locations and channels to communicate with potential customers. It is therefore an outcome of perceived relevance of the selected locations to the brands being sold. There has been no qualitative research that explores the perceived relevance of fantasy realms to marketing. Therefore this research aims to establish, through exploratory qualitative research, the perception of these locations in terms of constructed positioning. It asks if fantasy realms within immersive virtual environments (IVE's) are sacred space within which no marketing activity is possible without negative reactions and brand resistance taking place. As such the core contribution of this research is that through exploring perceptions of a number of fantasy realms within Second Life (SL), the literature on virtual environments, branding, and debranding, the nature of opportunity can be gauged.

II. Branding and Social Media

A key consideration was suggested by Ref. [15] regarding the rapid deployment of branding activity into social media and online communities as marketers are becoming confronted with the stark realization that social media was made for people, not for brands. The development of these IVEs has developed to a stage where they can realistically be called Virtual Worlds where the immersion and interaction develop in a world of their own. The authors further suggest that for brands to gain coveted resonance, the brand must relinquish control [15]. A further consideration for branding was proposed [25] that suggested that whilst branding aims to influence the complex social conditions for branding due to successful branding relying on patterns of social interaction that are not only beyond the control of brand managers, but cites the autonomy of such patterns of social interaction that are crucial to the authenticity of the brand. He further suggests that thinking in terms of place branding reveals attempts to force the brand on local actors [25]. According to Ref. [51] which suggests that the misuse of social networking sites may be defined as failure to engage sincerely. However, the current extent of any misuse in Immersive Virtual Social Environments is unclear.

Research into Perceived Relevance and Value of Advertising in Online Communities found that if online networking communities want to achieve positive member responses to advertising, they should consider two key perceptual factors that have pivotal effects on behavioral responses: perceived relevance and value of community advertising [45]. Specifically, when users perceive community advertising as more relevant to the theme of community and thus more congruent to the extension of their social identities, they regard that advertising as more valuable and exhibit more positive behavioral responses to it.' Ref. [52] suggest that 'Online social networking communities are digital networks in which users feel an intrinsic connection to other members.

According to Ref. [1] consumers typically utilise and transform mythic and symbolic resources within the marketplace to construct narratives of identities enriched with symbolic meanings. In this way the particpants develop a community. However, research suggests that some consumers define their identities through an ongoing opposition to cultural and lifestyle norms [1].

Groups, or communities, of such consumers may become more of a liability to the communication of mythic and symbolic brands and as such erode brand equity. Consumer ideologies of resistance may result in the questioning of corporate interests and motives of such brand communications. The framing of such a perspective may be through the creation of lifestyles and identities that ‘defy consumerist norms’ [1] and offer identification for a subject based upon belonging, triggered by available signs and representations within perceived spatial volatility that expresses direct experiences of identity [25].

Ref [21] suggested that in this area ‘identity formation as a process is spatially situated and thus one that is about creating symbolic spaces [21]. Whereby previous research has focused upon the anti brand movement [14] and consumer resistance and creativity [21], this research focuses upon the potential for brand resistance in immersive virtual environments such as Second Life. The perspective that all IVEs share the same views of resistance to commercial activities and exploitations in the market place of the real world may be too simplistic. However resistance to consumerist culture in some immersive virtual environments may share greater similarity to Ref. [12] work on resistance to consumer culture by lovers of the natural world. This may prove fruitful due to the type of experience realms [4] that exist in immersive virtual environments (IVEs) that may be better classified as ‘sacred space’. The 4th Strand social network [8] of Second life may provide an increasing number of implications for brand communications to be carefully executed.

One of the key elements in the interactive space of the virtual worlds is the lack of materiality; their virtuality. It has been argued that the materiality of interaction is a factor that has been neglected in the understanding of markets [46]. This direct material engagement with products is absent during the interaction that occurs in the virtual world. The virtual nature of this world tends to bring forward the phenomenological aspects of marketing and the development of intentional objects [50]. It will however important to consider the virtuality together with the construction of such worlds. In this situation where the consumer engagement through possesion of branded goods is unavailable the process of brand building process needs to be reconceptulized.

III. Branding and Brand Equity

According to Ref. [44] building brands involves a number of elements that includes product attributes – that are intrinsic distinguishing and non-distinguishing elements – and the non product ‘brand’ attributes that are intangible extrinsic elements [10]. It is these intangible extrinsic symbolic elements such as the logo, brand, trademark etc that underpin the basic contributor to developing associations to the brand and build recognition and awareness. As such it is these fundamental elements that are the focus of this paper. Branding, importantly, aims to balance social forces with communication strategies according to Ref. [36] but however often fails to recognize the potential and challenges of these social forces [25] Brand management can be conceptualized as the utilization or creation of space for the interaction of actors within social and cultural norms that is outside the brand owner organization that reduces the opportunity for resistance [25]. This is with a desired outcome of building brand equity.

Brand equity is defined as “the added value with which a brand endows a product” [41]. According to Ref. [29] brand equity lies in the opportunity space between three important components; value proposition, brand name (awareness) and product or service experience. Ref. [50] argues that ‘there are four dimensions of brand equity are brand awareness, brand associations, perceived quality, and brand loyalty’. It has been accepted for some time that brand awareness is one of best predictors of purchase’ [2].

Brand differentiation as “a clear performance differential over competition on factors that are important to the target customers” [29], is essential in being able to support premium pricing and creating a positive brand influence on buying behaviour [34]. For instance involving a Second Life island for a brand could help to build a brand community and enhance brand differentiation and image through professionalism and creativity. Internet practitioners consider brand equity important [27], however through technological enhancements the brand equity model needs to be developed to incorporate these developments in IVEs. The web equity framework by Ref. [41] is rooted in the original brand equity model, but involving dimensions of web awareness and web image that are defined as the consumer familiarity and perceptions about a Web site. Others incorporate this idea of community together with interactivity and value [47]. These models are very useful in aiding and building relationships, however there are limitations of in-depth detail to the resonance level, consequently it is difficult for the concept to create value to the consumer [42] and new media and technology information such as the virtual platforms are not integrated into the models.

The branding literature suggests explicitly how branding takes place in controlled spaces such as Nike towns, Prada boutiques or Starbucks coffee shops but offers little explanation of how branding influences or fails to influence outside these controlled spaces [25].

IV. Immersive Virtual Environments

Virtual worlds operate via Immersive virtual environment technology (IVET) allowing a consumer to feel more at one with his/her surroundings increasing the level of resonance felt. It has been identified that immersion in an IVE heightens the perceptive experience of individuals [3] [31].

The success of Virtual worlds is due in part to this sensation and occurrence of immersion. Ref. [5] suggests that full immersion can be likened to an altered state of consciousness claiming that you can daydream in a virtual world without leaving it. They present rich layers of synthesized sensory cues to the user so that they feel enclosed by the mediated environment and are willing to believe that the environment is real [52]. Virtual environments are technologically synthesized sensory information that makes the environment and their contents seem real [6]. The challenge lies in understanding the culture of virtual worlds and firms have to understand that this is the solution for being successful in these days: “If you love it, let it go.” [7].

Second Life was founded to target adults for socializing, however marketers have found many opportunities to enhance brand equity and differentiation to enhance loyalty within this virtual platform. Traditional channels for marketing are less noticed by consumers [9] as barriers have been reduced by new techniques and technologies, therefore brands have had to find new mediums to grab the attention, interact and build relationships with consumers, with the immersion into virtual worlds has been shown to aid in this situation.

Second Life residents visit this virtual world almost as if it were a real place, exploring what others have created, meeting other residents, socializing, exploring intimacy and love, participating in individual and group activities, and buying items (virtual property) and services from one another [39].

The virtual world ‘Second Life’ allows the user to express their imagination in teleporting to a wide array of fantasy islands. It offers a sense of escapism whilst remaining life like. There is considered to be a greater role for imagination in these IVE’s [11]. ‘Avatars lifelike behaviours can make a site more engaging and motivating, making customer interaction with the website much smoother’ [35]. Imaginative realms often need the protean nature of avatars. The relationship of play and construction is important in this virtual world. So these IVEs become life like although the separation from the real world is always acknowledged at a residual level. There is a constructed form of escapism and play. In a discussion of the form of play connected with the plastic arts [26] sees very little play in this art form. The static nature of this art can be compared with the performing arts and their construction of a play-sphere. Activity in these arts is often one of rapture and immersion. When considering the plastic arts such as sculpture, it is noted that ‘where there is no visible action there can be no play’ [26]. While there is a role for decoration and ritual in the engagement with the plastic arts what seems to be new here is that the plastic construction of virtual space requires immersive action with which to engage with it. So compared to traditional media, virtual environments better engage consumers and reproduce the real use experience [17], offering superior control and tailoring of messages [48].

In a realization of this role the focus of advertising in virtual environments is around interactivity and presence. Interactivity refers to a characteristic of a medium in which the user can influence the form and or content of the mediated experience [19]. The argument developed in this paper is that the concepts of presence and interactivity have a centrality in second life that relate to being positioned in the virtual space. Further developing this idea it can be seen that the form of constructed positioning is one that is particularly characteristic of second life. The software is designed to create an available space for interactivity. This interactivity is different in form that is generated from other technical environments such as websites and social media. The particular process of interactivity developed in this IVE depends on the requisitioning of the objects present in this virtual domain. We requisition them in order to make use of them; we interact with the other avatars that are present in the same virtual space. This interaction is generally with the objects that are virtual people. These avatars can be said to be requisitioned in a manner that relies on co-creation. The adoption of them in their interactivity requires greater immersive aspects than is the case in ‘real-world’ interaction. More specifically the immersive aspects are placed in the foreground to a greater extent by the assemblage created by this environment. These encounters with virtual objects require the process of selection from the created environment. This created environment can be seen as the standing reserve [18] the things from which we make a selection. The objects are called forward from the reserve in this way. This activity of making available can be seen as a crucial activity in Second Life; engaging in these activities provides the vitality and the interaction in the virtual world.

In what way does branding affect this calling-forward. In one way the brand communication provides ready recognition of those objects in the virtual world that are branded. The intention of branding in the literature is complex, its purpose is either to provide meaning [27] or, revealed in a discussion of corporate branding, to provide image [13]. Through both of these elements it can be said to promote connection or recognition in a potentially strange world. The branding by its very nature calls itself forward and demands availability. As branding is maintained over time there have been various strategies adopted towards branding in the virtual space.

V. Domains of Interest and Experience Realms

Much psychology and social-psychology supports the notion that people select their domains of interest based upon their values. As such many domains of interest exist from Art to Extreme sports and painting to snowboarding. In addition to which there are many domains such as sports, technology, travel, and fashion etc whereby those with congruent values select to spend their time engaging with others with similar values.

It is these domains that offer an opportunity for building brand equity through debranding in order to ensure non invasive exposure that does not compromise the overall experience of those immersed in the virtual environment.

This discussion question is assuming increasing importance. Virtual usage by firms and customers is increasing year by year [39]. In 2010 $1.6 billion was spent by Americans to buy virtual goods for Avatars in virtual worlds [11]. Whilst accepting this trend others imply that virtual worlds are a fad and therefore perception of marketing opportunity may be limited [47]. More organizations are finding the shared value [44] and potential [50] of having a presence in online communities to sell products, provide a service or accommodate events [20] as they aid in building brand equity, brand revitalization, and increasing commitment and loyalty. This is likely to result in greater consumer repeat purchases, more consumer information and positive word-of-mouth (WOM). In fact the WOM effects can mitigate the negative effects of advertising in online communities. This suggests that the idea of interaction is powerful in this context. [31]. In addition virtual communities allow for stronger relationship ties as they give the members a sense of attachment and purpose [44] as well as “meeting social and psychological needs that can be categorized as follows: information, relationship building, social identity/self-expression, helping others, enjoyment, status/influence and belonging”. This can be an advantage for marketers as their avatar within the community can act as an advocate influencing consumers’ opinions and behaviours [32] and [38].

Ref. [4] divided the experience realms that can be found in the virtual world environment into the following 4 categories:

1) Entertainment (passive absorbed). In a virtual world, this would include the consumption of media content, or of live content, such as viewing a stage performance in the virtual world, watching a movie on a screen in SL or listening to music or radio.

2) Education (active absorbed). Various examples in a virtual world environment include tutorials and online lectures. There are many examples of universities and other organizations that are now using Second Life for educational and teaching purposes.

3) Escapist (active immersion). For example, casinos, themed areas and 'sims' (i.e., 3-D virtual games within SL) all provide this kind of escapism. An example of a 'sim' would be a virtual world area with a gothic theme or a science fiction combat theme.

4) Esthetic (passive immersion). A typical virtual world example would be visiting a museum in SL such as the Second Life International Space Museum, Second Louvre or the Open Art Museum.

5) Sacred Space. A socially meaningful space which is based on engagements beyond that of consuming. For example rich social spaces to escape the real world such as virtual islands, parks, etc

Ref.[4] categories of experience realms does not explicitly suggest that overlaps may exist such that an escapist realm may also have esthetic qualities and immersion implications of both a passive and active nature. In essence they suggest that there are clear delineations between the types of realms. There may also be the implication that an organization seeking to create superior brand equity for competitive advantage would enter any of the first four realms above better than its competitors. For example, an education institution looking to offer students a better experience of learning through Second Life must do so by making this experience more absorbing and through utilizing some of the strengths of other realms make it actively immersive. The authors herein recognize that the above experience realms are a very useful starting point and with the addition of an experience realm to cater for spaces where negative brand perceptions are likely such as Sacred spaces. This certainly adds value in highlighting some potential brand issues for marketers. The experience realm of sacred space (5) may be added to guide marketers in their brand communications so as not to erode equity. The characteristic of such space may be rich in cultural values and perceptions that possibly distort or work against the brands identity.

In virtual reality many different types and combinations of experience realms are likely to exist with a variety of levels of brand interaction and it is through examination of these that one may be able to determine the receptivity towards brand exposure. A key question that emerges in relation to this research is what constitutes Sacred Space? And does it actually exist within virtual immersive environments? These are the focus of this paper.

VI. SOCIAL BENEFITS, LIABILITIES AND THE ROLE OF DEBRANDING

Some of the successes of social networks are based on the principles of replicating the experience of the 'real world'. However there are also many examples of social networks bashing brands as a result of their invasive techniques.

Previous work has studied individual and communities resisting market logic. Examples from the offline world include festival goers aiming to exist – even momentarily- outside of previously established commercial structures [34]. There are well documented consumer accounts exhibiting anti-brand attitudes [24]. Taking this to the online realm, anti-brand communities have been identified to form against specific brands and to gain social benefits by collectively pulling together in their commercial opposition towards these social network dynamics [23]. Brands have also been found to participate as 'uninvited' and to cause commercial reactions when deciding to enter primarily consumer led online spaces such as social media and video sharing websites [15]. Such anti-brand issues call into question the conditions under which commercial acts and branding might be acceptable within IVEs. We suggest that the process of debranding might be a relevant process to consider in terms of providing the grounds for such conditions.

Ref. [43] discussed 'debranding' as a no-name marketing strategy gaining acceptance in relation to branded products. In a more modern twist strongly established brands can be found to spatially debrand themselves in particular settings. Indicatively, instances of targeted elimination of the brand name is becoming a popular way for companies to differentiate themselves or extend their business in an effort to appear less corporate and more integrated with the social spaces and actors within them [37].

The range of organizations engaging in such activities is diverse: from Starbucks' '15th Ave Coffee & Tea' shops, with only 'Inspired by Starbucks' on them [37] to zoo animals being sponsored by well known brands such as Nescafe, but with the brand logo font changed and carved on a wooden plaque to signify a more natural feel (Athens Zoo, Greece). Consumers have also been found to engage in related practices, for example car enthusiasts 'debadging' their cars and rearranging the letters of the brand name in a playful and personal fashion [22].

The issue of commercial fit with online user activities (such as online gaming and the presence of arousing commercial billboards as investigated by Ref. [16]) has been documented. Online debranding, through the adjustment or temporary removal of previously ascribed brand layers (such as Nescafe's reconfigured type font for a particular space), might provide the conditions for non invasive commercial practices and the immersion sought in IVEs.

One visits the Zoo to see animals and yet commercial branding results in various animals being commercially sponsored to capitalize on the opportunity to build awareness and equity. However one may consider the domain of animals in Zoo's as not fair game for capitalizing commercially on their situation by including brand logo's, names, fonts etc. As such this branding activity could be seen as somewhat invasive and compromising the brand equity. However when considering debranding (offering a transparent version of the logo or standard font for the brand name) the cues and connections between these brands may result in the brand becoming strengthened with a greater potential for building brand equity.

In the process of debranding or debadging many would suggest that Nike was the first, but now the practice of eliminating the brand name from products or marketing activity is becoming a popular way for companies to differentiate themselves or extend their business [37]. The debranding has often occurred in a situation of familiarity, and the playful nature of debranding often relies on previously established recognition. The counterintuitive nature of dropping a well-known company name or logo from a product or marketing activity, offers a debranding strategy to make their companies and brands appear less corporate and more forward-thinking [36] and less separate from the social spaces and actors within them.

VII. METHODOLOGY

Exploratory qualitative research was undertaken with a cohort of 16 2nd year marketing students aged 19-22 who were selected as a result of their marketing knowledge regarding brands and branding and as they represent the future commercial marketing decision makers. They students were asked to explore a Second Life Fantasy Realm of their choosing. Then the students were interviewed to ascertain their views on the limitations and marketing opportunities that exist within that chosen realm. A total of 10 fantasy realms were explored.

VIII. FINDINGS

The fantasy realm locations selected by the research participants included, Winterfell x4, Dirty Talk Bar, Virtual Reality Project, Soul Sensation, Caribbean Resort, Eleven Caves Siden, the Doomed Ship x2, Java Island, Spuzikuzi, and Miami Beach.

A range of motives were perceived as the reasons why someone would visit the research participants chosen realms including; nostalgia x2, for adventure and fun, to see something different from the real world, looking for excitement and adventure, escapeism x2, visit somewhere you could not afford to go, relaxing, socializing, gain human interaction, exploring, couples experiences, for a beach experience, to find love, take part in sports, and enjoyment.

Those research participants that suggest that marketing would not be suitable for their chosen realm were for the following reasons;

The chosen realm was a Soul Sensation

"Too much commercialization would spoil the atmosphere" Emma.

The chosen realm of the Doomed ship

"You can't advertise using posters as it would ruin the environment" Evie.

The chosen realm was Java Island

"It is beautiful but not realistic. Due to the chilled out environment, I doubt that product placement would be valued in this realm" Maria.

The majority of research participants that suggest that marketing may be suitable was based on the following statements;

The chosen realm of Winterfell

"The buildings and realm as a whole is quite busy and full and if fully immersed they are likely to react positively to marketing that is blended in" Hannah.

"Maybe marketing would work in the houses i.e. sofa's etc so as not to conflict with the nature in Winterfell" Amanda.

"It is probably not relevant to contemporary / modern brands – but old world brands may fit with the theme" Andrew

The chosen realm of the Doomed Ship

"Everything is dark and spooky – so you could play with the design as long as it fits in with the theme such as horror movies" Louisa.

The chosen realm was Soul Sensation

"Leisure products could be marketed that fit in with the beach theme" Jessica.

The chosen realm was Java Island

"The design might need to be boring – you could have avatar travel agents" Sadie.

The chosen realm was the Virtual Reality Project

"Lots of white space and a choice of quirky objects to pursue – could be more immersive if you purchased something" Robert.

The chosen realm was the Dirty Talk Bar

"A stylish bar – therefore themed cocktails might work" Chris.

The chosen realm was the Caribbean Resort

"Any tourism holiday companies and water sports, pubs/clubs would work" James.

The chosen realm was Spuzikuzi

"You could have agents to chat to and to show off their art work" Janice.

IX. DISCUSSION: IVES: THE PRIMACY OF SACRED, YET COMMERCIALLY PERMEABLE, REALMS

The findings relate to the issue of immersion in online environments [5]. A key theme emerging from the data is the primacy of the online setting and the secondary - and complementary- role of any potential commercial activity. The articulated perceived reasons for a user to visit these online sites exhibited no explicit commercial activity as a perceived driver. Furthermore, there was a commonality of drawing on the importance of relevance and of preserving the 'environment', the 'atmosphere' and the 'theme' of the online settings in the light of any prospective commercial activity.

In this sense, such online realms were viewed as the type of realm theorized by Ref. [4] as sacred spaces, realms with no primary market function and of a social value ('for socializing, 'to gain human interaction, 'for couples' experiences', 'to find love' in our informant's words). Yet, although sacred to commercialization to a degree, they were still perceived as potentially permeable and open to marketing and branding activities under certain conditions. Such conditions, as explained above, were viewed around the need to follow the already established realm character, and resonate with it. They were also related to the concept of segregating spaces and finding permeable commercial spaces within the wider sacred realm: as noted for example branding activity could be kept within the materially based houses rather than in the wider natural realm (Winterfell).

X. Conclusions and implications for brand management in fantasy realms

The outcome of this conceptual paper involves the concept of relevance. Relevance is a key concept to consider in addition to differentiation. The Young and Rubicam brand asset valuator considers relevance to be an important concept in a brands future financial performance [40]. Whilst its application was largely derived for the physical and real world the concept of relevance may offer significant value in exploring different experience realms within the virtual world. The idea of relevance must be connected to the vituality of such IVEs and the idea of the 'standing reserve' of availability that provides the interactive experience. The need to engage and effectively communicate with potential customers may provide a juxtaposition of eroding brand perceptions if traditionally applied to the context of experience realms. In addition avoiding particular experience realms that may offer little relevance and availability limits the potential exposure. As such exploring these realms for what they mean and adapting the core brand through debadging may facilitate brand exposure in otherwise taboo sacred realms.

'The issue of brand building in virtual worlds is embryonic… It is likely to follow a similar learning curve to other new media, such as the Web and mobile telephony' [4]. The question becomes – is it going to be business as usual for brand development in virtual worlds?

The implications for building a brand presence that is not invasive within the realms of sacred space and brand communities, that helps to increase brand equity stems from the concept of debranding explored within this paper. Although there may be limited branding opportunity for utilizing non congruent sacred spaces and the resultant visual cues and semiotic effect of the brand greatly lessened, if the reduction of cues takes place some additional exposure and resonance may take place. The sacrifice need not be a lack of a deep understanding of the brand. Where [53] suggests that organizations and brands run the risk of making themselves look separate by not interacting with the space and the audiences, is going against transparency and oneness, a minor reduction of brand cues may overcome this. There has been some discussion of the idea of calling forward the objects in the virtual realm that are available for our interaction. The idea of branding, in either developing awareness or promoting its relevance in this domain challenges the unique element of immersion and calling forward that is characteristic in this virtual world. The raison d'etre in the Immersive Virtual Environments is the experience. It is this experience which needs to be respected and a strategy that attempts to colonize the virtual world as part of an existing brand strategy is one that might not be appropriate but also runs the risk of failing to grasp the potentiality for these types of immersive environments

XI. Future Research

The aim of this exploratory paper was to establish to which degree fantasy realms could be viewed as sacred space. The current findings both conceptualize and pinpoint the marketing potential to enter such realms. Future work should aim to look closer at the conditions for acceptance in different realms where existing brands already operate and to establish the need and the suggested forms of debranding strategies.

References

[1] Arnould, E. J., & Thompson, C. J. (2005). Consumer Culture Theory (CCT): Twenty Years of Research. Journal of Consumer Research, 31(4), 868-882

[2] Axelrod, J. N. (1968), "Advertising Measures that Predict Purchase," Journal of Advertising Research, 8, 3-17

[3] Blascovich, J., & Bailenson, J. N (2011, in press). Infinite Reality - Avatars, Eternal Life, New Worlds, and the Dawn of the Virtual Revolution. New York: William Morrow.

[4] Barnes, S., Mattsson, J., 2008, Brand Value in virtual Worlds: An axiological approach, Journal of Electronic Commerce Research, Vol.9, No.3

[5] Bartle, R. (2003), Designing Virtual Worlds, New Riders, Indianapolis, IN.

[6] Blascovich, J., Loomis, J., Beall, A. C., Swinth, K. R., Hoyt, C. L., & Bailenson, J. N. (2002). Immersive Virtual Environment Technology: Just Another Methodological Tool for Social Psychology?. Psychological Inquiry, 13(2), 146-149.

[7] Burke, P.; Darras, D.; Gerenski, K.; Gordon, M. E. (2007): The Virtual Brand Footprint: The Marketing Opportunity in Second Life,

[8] Cachia R., Compano, R., Da Costa, O. 2007 Grasping the potential of online social networks for foresight, Technological Forecasting and Social Change, Vol.74. PP1179-1203

[9] Chaffey, D. and Smith, P.R. (2008), E-marketing Excellence: Planning and Optimizing your Digital Marketing, 3rd ed., Butterworth-Heinemann, Amsterdam.

[10] Costa, R., Evangelista, S., (2008),"An AHP approach to assess brand intangible assets", Measuring Business Excellence, Vol. 12 No: 2 pp. 68-78

[11] Denegri-Knot, J., and Molesworth, M. (2010) 'Concepts and Practices of Digital Virtual Consumption', Consumption Markets and Culture Vol.13, No.2, pp109-132

[12] Dobsha, S., 1998, The lived experience of consumer rebellion against marketing, In Alba, J.W., & Hutchinson(eds), Advances in Consumer Research, Vol.25, pp91-97, Provo, UT.

[13] Fetscherin, M., & Usunier, J. (2012). Corporate branding: an interdisciplinary literature review. European Journal of Marketing, 46(5), 733-753

[14] Fischer, E. (2001). Rhetorics of Resistance, Discourses of Discontent. Advances In Consumer Research, 28(1), 123-124

[15] Fournier, S., Avery, J.(2011) Business Horizons, May, Vol. 54 Issue: Number 3 p193-207, 15p

[16] Greengard, S. (2011). Social games, virtual goods. Association for Computing Machinery: Communications of the ACM, Vol. 54 (4), 19.

[17] Grigorovici, Dan (2003), "Persuasive Effects of Presence in Immersive Virtual Environments," in Being There: Concepts, Effects and Measurements of User Presence in Synthetic Environments, Giuseppe Riva, Fabrizio Davide, and Wijnand A. Ijsselsteijn, eds., Amsterdam: IOS Press, 191-207,

[18] Heidegger, M., 2012, Bremen & Freiburg Lectures, Indiana University Press, Bloomington

[19] Heeter, C. (2000), "Interactivity in the context of designed experiences", Journal of Interactive Advertising, Vol. 1 No. 1,

[20] Hemp, P. (2006), "Avatar-based marketing", Harvard Business Review, June, pp. 48-57.

[21] Hetherington, K. (1998) Expressions of Identity: Space, Performance, Politics London, Thousand Oaks Ca., New Delhi: Sage/Theory, Culture and Society

[22] Hewer, P., & Brownlie, D. (2010). On market forces and adjustments: acknowledging consumer creativity through the aesthetics of 'debadging'. Journal Of Marketing Management, 26(5/6), 428-440

[23] Hollenbeck, C. R., & Zinkhan, G. M. (2006). Consumer activism on the internet: the role of anti-brand communities. Advances in Consumer Research, 33, 479.

[24] Holt, D. B. (2002). Why do brands cause trouble? A dialectical theory of consumer culture and branding. Journal of consumer research, 29(1), 70-90.

[25] Hornskov, S.B. (2007) On the management of authenticity: Culture in the place branding of Oresund., Place Branding and Public Diplomacy. Vo.3. No.4, pp317-331

[26] Huizinga, J. (1949) Homo Ludens: A study of the play-element in culture. Routledge and Kegan Paul, London

[27] Jevons, C., Abbott, M., & de Chernatony, L. (2005). Customer and brand manager perspectives on brand relationships: a conceptual framework. Journal of Product & Brand Management, 14(5), 300-309.

[28] Jobber, D. and Fahy, J. (2006), Foundations of Marketing, 2nd ed., McGraw Hill, New York, NY.

[29] Kapferer, Jean-Noël (2004), The New Strategic Brand Management: Creating and Sustaining Brand Equity Long Term, (Third ed.), London: Kogan Page.

[30] Keeling, K.A., P.J. McGoldrick, and S. Beatty. 2007. 'Virtual Onscreen Assistants: A Viable Strategy to Support Online Customer Relationship Building?' Advances in Consumer Research 34: 138-144.

[31] Keng,C-J., Ting,H-Y., Chen,Y-T., 2011, Effects of virtual-experience combinations on consumer-related "sense of virtual community, Internet Research, Vol. 21 Iss: 4 pp. 408 - 434

[32] Kim, J., Chio, J., Qualls, W. and Han, K. (2008), 'It Takes a Marketplace Community to Raise Brand Commitment: The Role of Online Communities', Journal of Marketing Management, 24(3-4), pp. 409-431.

[33] Kozinets, R. V. (2002). Can consumers escape the market? Emancipatory illuminations from burning man. Journal of Consumer Research, 29(1), 20-38.

[34] Kuhn, K.-A.L., Alpert, F. & Pope, N.K.L., 2008. An application of Keller's brand equity model in a B2B context. *Qualitative Market Research: An International Journal*, 11(1), pp.40-58.

[35] Luo, J.T. McGoldrick, P., Beatty, S., Keeling, K.A., (2006) "On-screen characters: their design and influence on consumer trust", Journal of Services Marketing, Vol. 20 Iss: 2, pp.112 - 124

[36] Lury, C. (2004) 'Brands: The Logos of the Global Economy', Routledge, Oxon, NY

[37] MarketingWeek (2012)Debranding the Great Name Dropping Gamble, 5th April. Accessed online http://www.marketingweek.co.uk/trends/debranding-the-great-name-dropping-gamble/4001018.article

[38] McGoldrick, P., Keeling, K. and Beatty, S. (2008), 'A Typology of Roles of Avatars in Online Retailing', Journal of Marketing Management, 24(3-4), pp. 433-461.

[39] Miano, T. J. (2007, August). Virtual World Taxation: Theories of Income. Retrieved March 2011, from Selected Work : http://works.bepress.com/cgi/viewcontent.cgi?.article=1000&context=timothy_miano

[40] Mizik, N., & Jacobson, R. (2008). The Financial Value Impact of Perceptual Brand Attributes. Journal Of Marketing Research (JMR), 45(1), 15-32.

[41] Page, C., Lepkowska-White, E, (2002) "Web equity: a framework for building consumer value in online companies", Journal of Consumer Marketing, Vol. 19 Iss: 3, pp.231 - 248

[42] Pappu, R., Quester, P.G. and Cooksey, R.W. (2005), "Consumer-based brand equity: improving the measurement empirical evidence", Journal of Product & Brand Management, Vol. 14 No. 3, pp. 143-54.

[43] Parasuraman, A. (1983). "Debranding": A Product Strategy With Profit Potential. Journal of Business Strategy, 4(1), 82-87.

[44] Porter, M. E., & Kramer, M. R. (2011). Creating shared value. Harvard Business Review, 89(1/2), 62-77

[45] Ridings, C. and Gefen, D. (2004), "Virtual Community Attraction: Why People Hang Out Online," Journal of Computer-Mediated Communication, 10 (1), http://jcmc.indiana.edu/vol10/issue1/ridings_gefen.html(accessed August 29, 2008).

[46] Shove, E., Araujo, L., 2010, Consumption, materiality, and markets, in Reconnecting Marketing to Markets, OUP, Oxford.

[47] Solomon, M.R., Bamossy, G. J., Askegaard, S. T., and Hogg, M.K. (2013) Consumer Behaviour A European Perspective 5th Edition, Pearson, London

[48] Tam K.Y., Ho S.Y., (2006). Understanding the impact of web personalization on user information processing and decision outcomes. MIS Quarterly, 30(4), 865-890.

[49] Tikkanen,H., Hietanen, J., Henttonen,T., Rokka, J., (2009),"Exploring virtual worlds: success factors in virtual world marketing", Management Decision, Vol. 47 Iss: 8 pp. 1357 - 1381

[50] Washburn, J. H., Plank, R. E. (2002). Measuring Brand Equity: An Evaluation of A Consumer-Based Brand Equity Scale. Journal of Marketing Theory & Practice, 10(1), 46.

[51] Wilson, A., (2012), What can phenomenology offer the consumer?: Marketing research as philosophical, method conceptual, Qualitative Market Research: An International Journal, Vol. 15 No.3 pp. 230 - 241

[52] Wellman, B and Gulia, M (1999), "Net-Surfers Don't Ride Alone: Virtual Communities as Communities," in Networks in the Global Village, B. Wellman, ed. Boulder, CO: Westview Press, pp 331-366.

[53] Witmer, B.G. Singer, M.J. (1998). Measuring Presence in Virtual Environments: A Presence Questionnaire. Presence, Vol. 7, No. 3, pp.225-240

[54] Yan, J. (2011) Social media in branding: fulfilling a need. Journal of Brand Management. Vol.18, No.9. pp688-696

Cross Language Information Retrieval Model For Discovering WSDL Documents Using Arabic Language Query

Prof. Dr.Torkey I.Sultan
Information Systems Department,
Faculty of Computers & Information
Helwan University,
Cairo,Egypt

Dr. Ayman E. Khedr
Information Systems Department,
Faculty of Computers & Information
Helwan University,
Cairo,Egypt

Fahad Kamal Alsheref
Management Information Systems
department, Modern Academy,
Cairo,Egypt

***Abstract*—Web service discovery is the process of finding a suitable Web service for a given user's query through analyzing the web service's WSDL content and finding the best match for the user's query. The service query should be written in the same language of the WSDL, for example English. Cross Language Information Retrieval techniques does not exist in the web service discovery process. The absence of CLIR methods limits the search language to the English language keywords only, which raises the following question "How do people that do not know the English Language find a web service, This paper proposes the application of CLIR techniques and IR methods to support Bilingual Web service discovery process the second language that proposed here is Arabic. Text mining techniques were applied on WSDL content and user's query to be ready for CLIR methods. The proposed model was tested on a curated catalogue of Life Science Web Services http://www.biocatalogue.org/ and used for solving the research problem with 99.87 % accuracy and 95.06 precision**

Keywords—Web services discovery; Cross Language Information Retrieval; WSDL; Text Mining

I. Introduction

Web service discovery aims at finding services whose description matches that of a desired service. The description of service contains a functional and a non-functional part. The former provides information about what the service does and how it works. This is basically expressed in terms of the required inputs and generated outputs, as well as any pre-conditions that need to be satisfied in order for the service to be executed and any effects that result from its execution. [9]

The discovery process is done through applying information retrieval techniques on the WSDL content of the web service. Large number of studies used various Information Retrieval (IR) techniques to search textual service metadata for service discovery. Wang and Stroulia [17] employed the inverted file [11] to index and search natural language description of desired services. Also Platzer and Dustdar [3] used the Vector Space Model (VSM) to implement a search engine for a Web service, and Sajjanhar et al. [1] have attempted to leverage Latent Semantic Analysis (LSA), a variant of VSM, for Web services discovery.

All of the above work index service metadata found either in a WSDL file (i.e., the <documentation/> tag) or from a Universal, Description, Discovery, and Integration (UDDI) registry entry. In both cases, service metadata written in English are manually created by service providers.

When user wants to search for a web service for a specific purpose he should write his query in English language because the service metadata is written in English language. The motivation for addressing this idea came from our experience in developing the web service and studying the Cross-Language Information Retrieval (CLIR) methods and data mining techniques. We deal with the web services as collections of documents that should be prepared for IR techniques, then we applied CLIR techniques to find the suitable service that matches the user query that written in other language. The process that is responsible for finding the suitable service is matchmaking process.it is the process of finding suitable services given by the providers for the service requests of consumers. The current service discovery mechanism of WSs is based on WSDL [8] and UDDI [2]. WSDL is an XML based language to describe properties of services that written in English language. UDDI is a registry where service providers can advertise their services and service consumers can search for services. The specific objective of our research is therefore to apply CLIR in the web services discovery; this is done by modifying the Match Maker process by adding CLIR components to support the Cross language web service discovery.

Applying this approach leads to enable different stakeholders to search for web services using their natural language, especially in the new operating systems like Android , Windows 8 metro applications and Apple IOS. The developers of applications that working on these platforms are interacting with web services to expand the application capabilities like connecting to database server, or generating reports from multiple database resources. Searching and finding such a service to perform the developer required function are the heart of our research, and as we mentioned before the search process was being performing by English language only, but in our proposed model we expanding the research process to let the service searchers to find their required service but with any language not only with English language. Here we proposed the Arabic language because in Arab countries we have a lot of good developers but without a good awareness with English language, so our proposed model should support developers to find and understand the web service description file which written in other languages. [12].

The rest of this paper is organized as follows. Section 2 provides the technical background that is used in the rest of the paper. Section 3 provides the related works that used in our research. Section 4 defines our proposed model and the modifications that had been done on the original one. Section 5 studies our methods in a case study. Finally, Section 6 contains the conclusion and future work.

II. Background And Motivation

In this section we provide a succinct background of the used techniques in our research. These techniques are: 1) Cross Language Information Retrieval, 2) Inverted file indexes, 3) WSDL term extraction, 4) WSDL term tokenization. And 5) Bilingual dictionary.

A. CLIR

Cross-language information retrieval (CLIR) systems allow users to find documents written in different languages from that of their query. The goal of a CLIR system is to help searchers find documents that are written in languages that are different from the language in which their query is expressed. This can be done by constructing a mapping between the query and document languages, or by mapping both the query and document representations into some third feature space. The first approach is often referred to as ‘‘query translation’’ if done at query time and as ‘‘document translation’’ if done at indexing time, but in practice both approaches require that document-language evidence be used to compute query-language term weights that can then be combined as if the documents had been written in the query language.[13]

In all cases, however, a key element is the mechanism to map between languages. This translation knowledge can be encoded in different forms as a data structure of query and document-language term correspondences in a machine-readable dictionary or as an algorithm, such as a machine translation or machine transliteration system. Gina-Anne Levow (2004) proposed CLIR reference architecture that focuses on the query translation architecture that illustrates the full range of opportunities to improve retrieval effectiveness. Fig. 1 present the Gina CLIR model that illustrates the data flow between the key components in her reference architecture. The dictionary based query translation architecture consists of two streams of processing, for the query and documents. Specifically, she exploits methods for suitable term extraction and pseudo-relevance feedback expansion at three main points in the retrieval architecture: before document indexing, before query translation, and after query translation. [10]

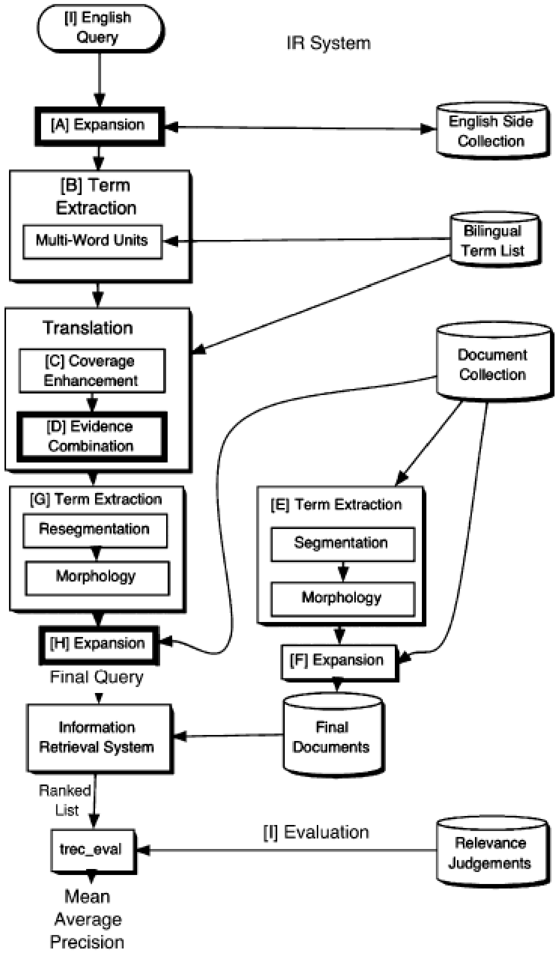


Fig. 1. CLIR architecture [10]

The previous architecture would be modified to accept queries written with other language like Arabic Indian etc. and the bilingual dictionary would be modified based the required languages for example Arabic vs. English Term list. The previous steps will be modified and explained in details in section 4.

B. Inverted file indexes

Inverted file is widely used for indexing text database. To support efficient information retrieval, the words of interest in the text are sorted alphabetically. For each word, the inverted file records a list of identifiers of the documents containing that particular term. Consider a sample text database consists of five documents. The indexer parses these five documents, and produces a set of distinct words for constructing the inverted file. The inverted file has two components a vocabulary and a set of inverted lists. The vocabulary comprises a collection of distinct words extracted from the text database. For each word t, the vocabulary also records: (1) the number (ft) of documents that contain t, and (2) the pointer to the corresponding inverted list. [15]

Each one of the word-specific inverted lists records: (1) a sequence of documents that contain t (notice that each document is represented as a document number d), and (2) for each document d, the frequency (fd,t) of t appearing in d. Thus, the inverted list is a list of < d, fd,t > pairs.[15]

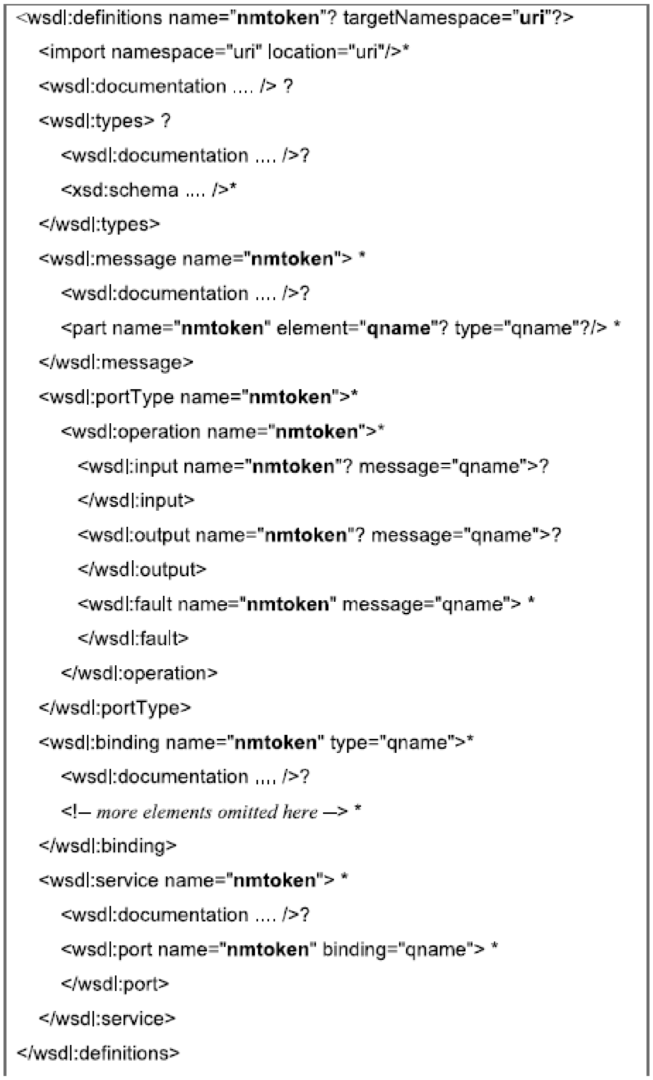
Applying the inverted file in WSDL discovery process will be done through dealing with WSDL files as a documents that containing the terms that needed to calculate the (fd,t) pairs the frequency (fd,t) of t appearing in WSDL document.

C. WSDL term extraction

WSDL document is an XML document validated against the WSDL schema. A lot of studies [16] in the XML retrieval literature suggested extract content from XML markups before applying IR techniques. Because the name attribute of an XML element represents the semantics for that particular element. Therefore, the value (i.e. nmtoken) of attribute name depicted in Fig. 2 is a good candidate for content extraction. [5]

Fig 3 presents a WSDL document instance. The value (e.g. GetLastTradePrice) of the name attribute for element <operation /> carries useful information implying the purpose of this operation at the lexical level. Similarly, values (i.e. nmtoken) of attributes ‘‘name’’ are extracted from a number of important WSDL elements (marked as bold faces in Fig. 3) definitions, message, part, portType, operation, input, output, service, and port. The value (i.e. qname) of attribute element for element part is also extracted for capturing the data structure of the parameters sent to/from the service operations. This forms recursive extraction of underlying data types for this element and/or type. In this example, the value (i.e. body) of attribute name for element part gives little useful information representing the real meaning of the input message part. Nevertheless, values (i.e. TradePriceRequest and TradePrice) of attribute element provide very valuable data in understanding the meaning of two message parts. The data structure within the WSDL element types and schema reveals more important lexical information about these two message parts through extracting the value of attribute name for element. Thus, in the case of TradePriceRequest, the element value is tickerSymbol. For TradePrice, price is extracted. Thus, by exploring solely lexical information, one can speculate that thisWebservice takes as input the stock ‘‘ticker symbol’’, and returns as output the ‘‘price’’ of the corresponding stock. The related operation name GetLastTradePrice also supports this proposition.[5]

The <documentation/> elements contain natural language information that could be used for constructing the text database.



```
<wsdl:definitions name="nmtoken"? targetNamespace="uri"?>
   <import namespace="uri" location="uri"/>*
   <wsdl:documentation .... /> ?
   <wsdl:types> ?
      <wsdl:documentation .... />?
      <xsd:schema .... />*
   </wsdl:types>
   <wsdl:message name="nmtoken"> *
      <wsdl:documentation .... />?
      <part name="nmtoken" element="qname"? type="qname"?/> *
   </wsdl:message>
   <wsdl:portType name="nmtoken">*
      <wsdl:operation name="nmtoken">*
         <wsdl:input name="nmtoken"? message="qname">?
         </wsdl:input>
         <wsdl:output name="nmtoken"? message="qname">?
         </wsdl:output>
         <wsdl:fault name="nmtoken" message="qname"> *
         </wsdl:fault>
      </wsdl:operation>
   </wsdl:portType>
   <wsdl:binding name="nmtoken" type="qname">*
      <wsdl:documentation .... />?
      <!— more elements omitted here —> *
   </wsdl:binding>
   <wsdl:service name="nmtoken"> *
      <wsdl:documentation .... />?
      <wsdl:port name="nmtoken" binding="qname"> *
      </wsdl:port>
   </wsdl:service>
</wsdl:definitions>
```

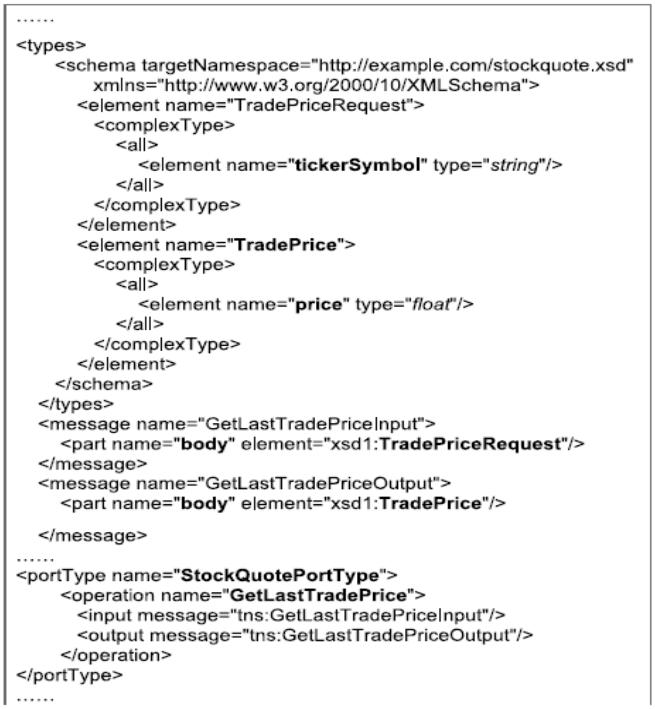
Fig. 2. WSDL document structure [9][5].

However, the reliability of comments or documentation written by humans is concerning as they may be either misleading or obsolete from the actual WSDL interface. So the WSDL parser is responsible for extracting the textual content directly from the interface definition to support information retrieval tasks.

D. WSDL term tokenization

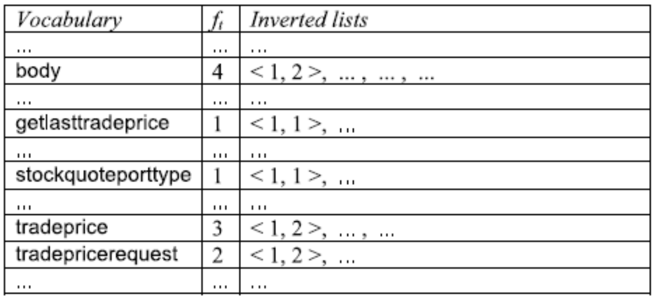
An example of WSDL document in fig 3 according to the WSDL document structure in fig 2.This document uses untokenized words and sentences to define WSDL element attributes such as names (bold face fonts). This is due in part to the rules defined in the WSDL standard: all WSDL element names are associated with the W3C Schema attribute data type ‘NMTOKEN’, which is a mixture of name characters (including letters, digits, combining chars, etc.) but excludes the single white space (#x20) [5][4]. Furthermore, in practice, most WSDL documents are not created directly by humans but are automatically converted from other high level programming languages (e.g. Java, C#, etc.), in which white spaces are prohibited in variable names. Consider the operation name GetLastTradePrice in Fig. 5. Under the normal text operation, the inverted file will create a new entry for the single word GetLastTradePrice in the vocabulary. [5]

As a result, the inverted file would be something looks like in Fig. 4, in which a partial vocabulary and its associated document frequencies ft and inverted lists (assuming the document id for this WSDL file is ‘1’) are presented.[5]



```
……
<types>
    <schema targetNamespace="http://example.com/stockquote.xsd"
         xmlns="http://www.w3.org/2000/10/XMLSchema">
       <element name="TradePriceRequest">
          <complexType>
             <all>
                <element name="tickerSymbol" type="string"/>
             </all>
          </complexType>
       </element>
       <element name="TradePrice">
          <complexType>
             <all>
                <element name="price" type="float"/>
             </all>
          </complexType>
       </element>
    </schema>
  </types>
  <message name="GetLastTradePriceInput">
     <part name="body" element="xsd1:TradePriceRequest"/>
  </message>
  <message name="GetLastTradePriceOutput">
     <part name="body" element="xsd1:TradePrice"/>

  </message>
……
<portType name="StockQuotePortType">
      <operation name="GetLastTradePrice">
         <input message="tns:GetLastTradePriceInput"/>
         <output message="tns:GetLastTradePriceOutput"/>
      </operation>
</portType>
……
```

Fig. 3. A sample of a partial WSDL document, source [9][5].



Vocabulary	f_t	*Inverted lists*
...	...	...
body	4	< 1, 2 >, ... , ... , ...
...	...	...
getlasttradeprice	1	< 1, 1 >, ...
...	...	...
stockquoteporttype	1	< 1, 1 >, ...
...	...	...
tradeprice	3	< 1, 2 >, ... , ...
tradepricerequest	2	< 1, 2 >, ...
...	...	...

Fig. 4. A sample of a partial WSDL document, source [9][5].

This should affect the IR process so the operation name has to be tokenized into four separate terms “GetLastTradePrice” to Get, Last, Trade, and Price. One may argue that a trivial string pattern discovery (e.g. letter case patterns) would easily tokenize them into meaningful terms. This Tokenization problem was solved and the solution is presented in the related work section.

E. Bilingual dictionary

A bilingual dictionary or translation dictionary is a specialized dictionary used to translate words or phrases from one language to another. Bilingual dictionaries can be unidirectional, meaning that they list the meanings of words of one language in another, or can be bidirectional [disambiguation needed], allowing translation to and from both languages. Bidirectional bilingual dictionaries usually consist of two sections, each listing words and phrases of one language alphabetically along with their translation. In addition to the translation, a bilingual dictionary usually indicates the part of speech, gender, verb type, declension model and other grammatical clues to help a non-native speaker use the word. Other features sometimes present in bilingual dictionaries are lists of phrases, usage and style guides, verb tables, maps and grammar references. In contrast to the bilingual dictionary, a monolingual dictionary defines words and phrases instead of translating them. [6]

Bilingual term lists are extensively used as a resource for dictionary-based Cross-Language Information Retrieval (CLIR), in which the goal is to find documents written in one natural language based on queries that are expressed in another.

The translation component of dictionary-based CLIR techniques depend on a successful cascade of three processes: (1) selection of the terms to be translated, (2) generation of a set of candidate translations, and (3) use of that set of candidate translations in the retrieval process. For the first stage, the best results are typically obtained by translating multiword expressions when possible, backing off to individual words when necessary, and further backing off to morphological roots when the surface form cannot be found [6].

In the second stage, algorithms for choosing among alternative translations have been extensively studied, and older techniques based on averaging weights computed for each translation can benefit significantly from translation selection based on term co-occurrence within the target corpora. The focus on the third stage has been somewhat more recent, with the best presently known technique based on accumulating term frequency and document frequency evidence separately in the document language, then combining that evidence to create query-language term weights [6].

III. RELATED WORK

A. The IR-style Web Services Discovery

Numerous recent efforts have been reported in applying IR for Web services discovery. Here we proposed Chen’s IR-style web service discovery that uses inverted file indexes for Web services discovery that will be merged in CLIR methods in our proposed approach.

Chen Wu proposed IR-style Web services discovery approach that was illustrated in Fig. 5, in which term tokenization constitutes an important step for WSDL term processing. Initially, service providers deploy their Web services accessible to the public via the Web. In doing so, they also publish a service description, i.e., the WSDL documents, which captures the functional capabilities and technical details (e.g., transport bindings) of a Web service.

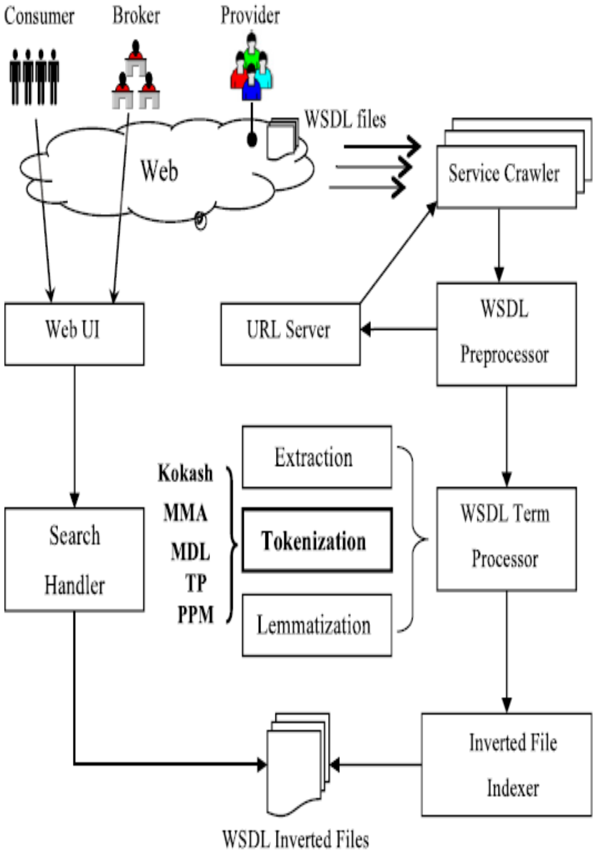


Fig. 5. IR-style service discovery approach. [5]

These service descriptions can be collected by a number of Service Crawlers, which fetch WSDL files from the Internet. Alternatively, they can also be collected from some well-known service datasets such as one of the WS curated catalogue .Crawlers hand over retrieved WSDL files and associated HTML files to the WSDL Preprocessor for link analysis. This yields a list of new URLs that may point to some new WSDL files. These URLs are assigned to an idle crawler by the URL server such as Pica-Pica Web Service Description Crawler. [5]

All retrieved WSDL files are then passed to the WSDL Term Processor, which (1) parses WSDL files and extracts important data (e.g., operations, messages, data types, etc.), (2) tokenizes extracted content into separate terms (the focus of this paper), and (3) carries out other linguistic tasks such as lemmatization and stop-word elimination, etc. Five tokenization methods (two baselines – Kokash and MMA – and three statistical methods MDL, TP, and PPM) are used in (2).

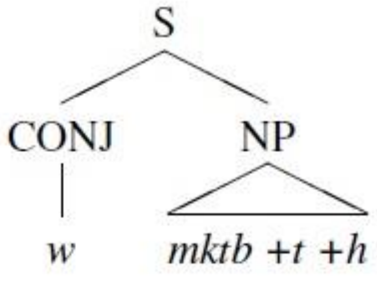
The WSDL Processor generates the 'term document', which contains separated words in a flat structure. The term document is transferred to the Inverted File Indexer. The indexer takes as inputs tokenized and lemmatized terms with their associated occurrences information in each document and generates as outputs the compiled data arrangement with pre-aggregated information optimized for fast searching. The data structure of inverted index is consistent with the notion of term-document matrix, which consists of term vectors as matrix rows and document vectors as matrix columns. The term vector is sorted that allows fast lookup operation. After finishing the document tokenization and indexing processes, the search handler component will extract the key terms from the query then it will search via extracted terms in the inverted file index , after that the most matched WSDL documents is returned based on terms frequency in the WSDL documents [5].

But Chen's model is based on the document retrieval where the query and the documents are written in the same language ,but if the query was written in different language it will not retrieve the documents or the services, for this limitation the need for applying the CLIR techniques was arising to support searching in documents with queries with different languages, so we modified Chen's model to accept queries with different languages "Arabic in our model" and retrieve the suitable service and WSDL document with a translated WSDL version to query writer's language.

B. *Arabic Treebank (ATB) segmentation*

The Arabic language has a very rich morphology where a word is composed of zero or more prefixes, a stem and zero or more suffixes. This makes Arabic data sparse compared to other languages, such as English, and consequently word segmentation becomes very important for many Natural Language Processing tasks that deal with the Arabic language

The ATB is used in most of Arabic language segmentation cases, this is a light segmentation adopted to build parse trees in the Arabic TreeBank (ATB) corpus. This type of segmentation considers splitting the word into affixes if and only if it projects an independent phrasal constituent in the parse tree. As an example, in the word ومكتبته (wmktbth — and his library) mentioned earlier, the phrasal independent constituents are: (i) conjunction و (w — and); (ii) nounand the head of a Noun Phrase (NP) مكتبة (mktbt library); and (iii) a pronoun (PRON) ه (h—his). This would lead to the following parse tree: [14]



A full segmentation (i.e., morphological segmentation) will separate the suffix ة(t feminine marker) from the word **مكتبة** (mktbt -library).Since the ة (and generally all the suffixes which are gender marks) are not independent constituents as shown in the previous parse tree, they are not considered for ATB segmentation. Thus, the ATB segmentation scheme considers splitting only a subset of prefixes and suffixes from the stem. When using ATB segmentation, the number of words is similar to its counterpart in English. This is one reason why

ATB segmentation is widely used in building machine translation systems for the English-Arabic language pair. For the word ومكتبته (wmktbth - and his library), the ATB segmentation would be و + مكتبت +ه (w+ mktbt +h). Prefixes that are considered for possible segmentation are:

1. “ل”, (l — to);
2. "ب" (b — in);
3. “و” (w —and); and
4. “ك” (k — as).

Possible segmented suffixes are the possessive personal pronouns such as:

1. “ي” (y— my);
2. “هم” (hm — their);

F. “كم” (km— yours); etc.[14]

C. Approaches to CLIR

The current CLIR research into three categories [7]:

- Document Translation.

Document translation comprises approaches to CLIR which require that all documents in the collection be translated into the language of the original user request. User requests or derived queries are then dealt with by a monolingual IR system.

The principal translation method reported in the literature is commercial off the shelf MT. The rationale behind this approach is that whereas user requests are often viewed as being too short to provide sufficient contextual information for traditional MT to perform well, documents may be translated as normal texts. This approach may be implemented without developing any CLIR-specific software. Nothing is needed other than a commercial MT product and a standard retrieval engine. Problems such as translation ambiguity and coverage are dealt with in a "black box" manner by the MT software.

There are several problems with document translation for CLIR. The first is that the cost in terms of time and money of translating a large document collection using traditional MT technology can be prohibitive. It took Oard and Dorr two months to translate the 550MB TREC German collection.

- Non-Translation Based Methods.

There is a small number of approaches to CLIR which translate neither the requests/queries nor the documents, opting instead to convert both to a language-independent representation where they can be searched directly .The only such system to be entered in a large-scale evaluation such as TREC was the CINDOR system at Textwise Corporation which used Wordnet synsets as a multilingual thesaurus to mediate between requests and documents. However, despite the existence of projects like EuroWordNet which aim to translate Wordnet into languages other than English by hand, Wordnet is still limited in its coverage, and it is difficult to see how it could be expanded without considerable work. Considerable improvements in performance were recorded at TREC-8 for the CINDOR system by switching to using MT software for request translation.

- Request or Query Translation

We have seen that it is not usually feasible to translate each document in the collection into every language represented in it, and that existing techniques which map both documents and queries or requests to an Interlingua representation require as much hand-crafted knowledge as document MT but do not perform as well. In this section we examine the obvious alternative - translating the requests or queries into the languages of the document collection. There are three main query translation methods:

- Request MT. This is where a commercially-available MT engine is used as a "black box" to translate the user request as-is.
- Corpus-Based Query Translation. This is where techniques from the domain of corpus linguistics are applied to map the terms in the bag of words query derived from the user request to a semantically equivalent representation in the target language.
- Dictionary-based Query Translation. This is where a simple machine-readable bilingual dictionary is employed to map the terms in the bag of words query to an equivalent representation in the target language.

IV. ARABIC CLIR WEB SERVICE DISCOVERY MODEL

A. The Arabic CLIR Web services discovery approach is illustrated in fig 6.

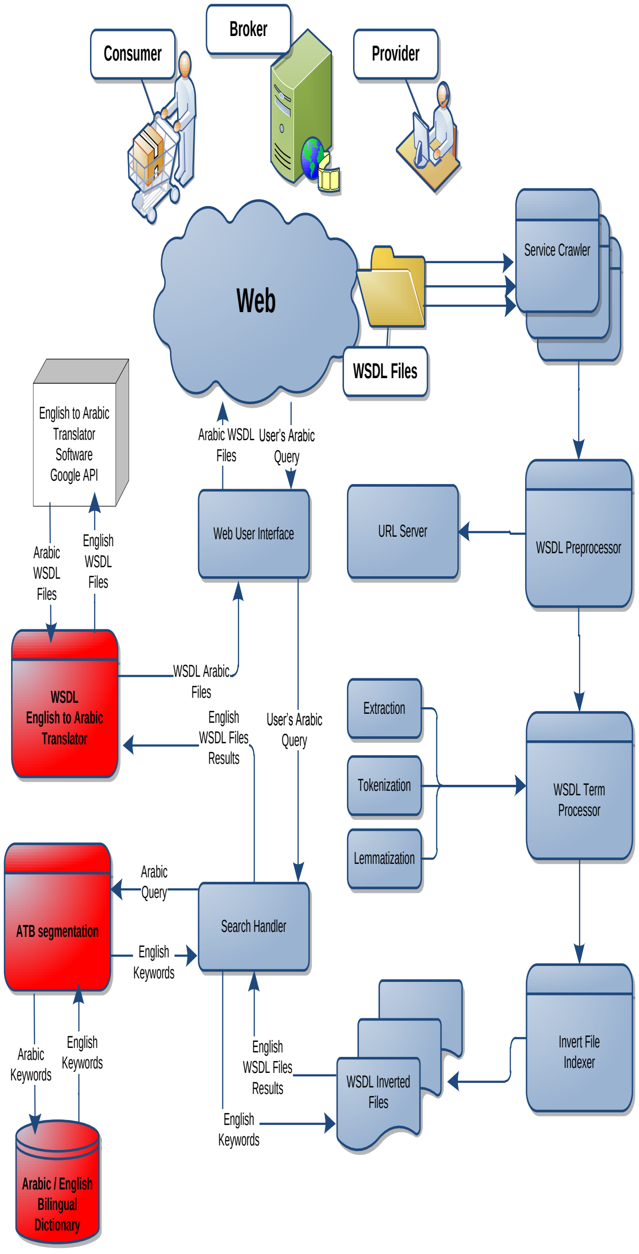


Fig. 6. Cross Language Information Retrieval service discovery approach.

As we mentioned in the IR-web service discovery in the related work part, the Service Crawler is responsible for collecting the WSDL document for the services, each service has its own WSDL file that was written by the service provider.

The WSDL preprocessor and WSDL are responsible for preparing the WSDL files and extract nmtoken values that mentioned in the WSDL tokenization part in this paper. All WSDL files are transferred to WSDL inverted files by extracting the WSDL nmtokens values and put them into the same form as fig4.

Our contribution is presented in adding the CLIR components to IR-web service discovery approach. This components will expand the WS discovery to other languages like Arabic and etc.in our approach we are working in Arabic language only.

B. CLIR Components

The user will send his query through the web user interface to the search handler component, this query may be written in English language or in Arabic language, and if the query was written in English then the search handler will search inside WSDL inverted files without any change in the IR-web service discovery.

But if user writes his query in Arabic then the search handler will use our CLIR components the detailed in the following sections.

C. ATB Segmentation

We chose the Query Translation approach to apply CLIR. The query translation approach is cost efficient especially if it was compared with the translation of each document in the collection into every language represented in it.

The first step in the Arabic Query Translator is query segmentation. We used The ATB segmentation algorithm for divide the user's Arabic query into several tokens .as we mentioned before number of Arabic tokens is similar to the English one which makes the translation process is easier. The following example shows the Arabic query as an input and the segmented Arabic tokens as an output of the process.

Arabic query:

تحويل من PDF الى نص

After applying the ATB:

(S (NP <tahweel تحويل

(PP <min من

(NP< PDF **PDF**

(PP <ilaY إلى

(NP< nas نص

)))))

After applying the ATB segmentation algorithm the output is a several tokens which each one could be an entry in bilingual dictionary in the next step.

D. Arabic / English Bilingual Dictionary

As we mentioned before the bilingual dictionary is a specialized dictionary used to translate words or phrases from one language to another. For this purpose we searched for Arabic / English open source dictionary. We found ARABEYES, it is a Meta project that is aimed at fully supporting the Arabic. It is designed to be a central location to standardize the Arabization process.

The extracted tokens from the ATB are sent it to the ARABEYES to get the translated keywords as shown in the following example:

Arabic Tokens:

- **تحويل**
- **من**
- **PDF**
- **إلى**
- **نص**

ARABEYES output:

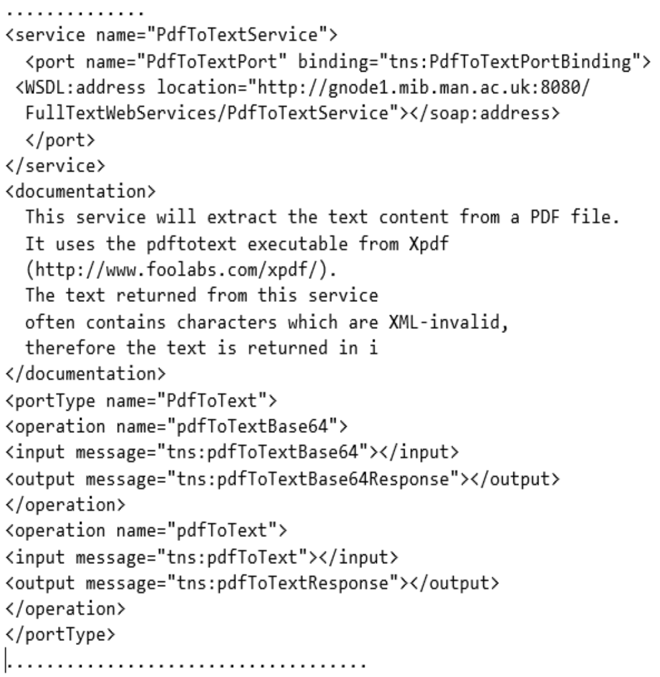
- **Conversion تحويل**
- **From من**
- **PDF PDF**
- **To إلى**
- **Text نص**

The Translated keywords are sent to the search handler components. Search handler components are using the inverted file techniques to find the WSDL documents that matches the user's query keywords. Each one of the translated keywords is compared to inverted lists records where each WSDL documents that contain t (notice that each document is represented as pairs < d, fd,t > d is a document number and the frequency (fd,t) of t appeared in d. then the arranged WSDL document list according the term frequency in the document is returned to the search handler component.

E. *WSDL English to Arabic Translator*

The search handler component gets the WSDL Documents that matching the user's query. All these documents are written in English language. These document should be translated to Arabic, before the translation process the information inside the WSDL document should be extracted. As we mentioned in section 2.3 the important information is exists in the attribute names that contain a semantic information about the selected web services and its values contain the information that is related to the previous attributes, the other important part in the WSDL document is the <documentation> element. This part is written in natural language as a description for the web service, which describe the purpose of the service and other information provided by the service provider, the following example in fig7 shows some of this information:

Attributes and its content:



```
..............
<service name="PdfToTextService">
  <port name="PdfToTextPort" binding="tns:PdfToTextPortBinding">
 <WSDL:address location="http://gnode1.mib.man.ac.uk:8080/
  FullTextWebServices/PdfToTextService"></soap:address>
  </port>
</service>
<documentation>
  This service will extract the text content from a PDF file.
  It uses the pdftotext executable from Xpdf
  (http://www.foolabs.com/xpdf/).
  The text returned from this service
  often contains characters which are XML-invalid,
  therefore the text is returned in i
</documentation>
<portType name="PdfToText">
<operation name="pdfToTextBase64">
<input message="tns:pdfToTextBase64"></input>
<output message="tns:pdfToTextBase64Response"></output>
</operation>
<operation name="pdfToText">
<input message="tns:pdfToText"></input>
<output message="tns:pdfToTextResponse"></output>
</operation>
</portType>
|.................................
```

Fig. 7. Example of WSDL document returned form the search handler.

The extracted WSDL documents are sent to the WSDL English to Arabic Translator component. The attributes names and its values are translated to Arabic. For this purpose they will be sent to WSDL translator component as a pairs like in the following example:

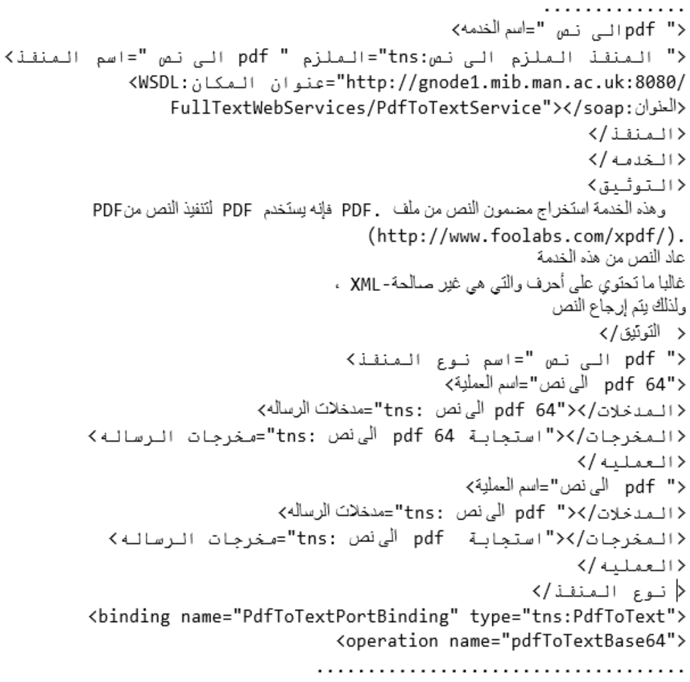
< (Attribute Name) service name=
(Attribute Value)"PdfToTextService">

And the output will be like:

<اسم الخدمه=" الى نصpdf ">

After translating all WSDL attributes and their values the WSDL translating components will collect all translating process output and put them in the form like figure 8.

The two WSDL documents the English and the translated are sent to the user or to the service requester, which will give him a better understanding of the WSDL documents, and let him to decide which service is the best match for his query.



```
...............
<" pdf الى نص "=اسم الخدمه>
<" المنفذ الملزم الى نص:tns"=الملزم " pdf الى نص "=اسم المنفذ>
<WSDL:المكان عنوان="http://gnode1.mib.man.ac.uk:8080/
FullTextWebServices/PdfToTextService"></soap:العنوان>
</المنفذ>
</الخدمه>
<التوثيق>
وهذه الخدمة استخراج مضمون النص من ملف .PDF فإنه يستخدم PDF لتنفيذ النص منPDF
(http://www.foolabs.com/xpdf/).
عاد النص من هذه الخدمة
غالبا ما تحتوي على أحرف والتي هي غير صالحة- XML ،
ولذلك يتم إرجاع النص
</التوثيق >
<" pdf الى نص "=اسم نوع المنفذ>
<"pdf 64 الى نص"=اسم العملية>
<المدخلات/><"pdf 64 الى نص :tns"=مدخلات الرساله>
<المخرجات/><"استجابة pdf 64 الى نص :tns"=مخرجات الرساله>
</العمليه>
<" pdf الى نص"=اسم العملية>
<المدخلات/><" pdf الى نص :tns"=مدخلات الرساله>
<المخرجات/><"استجابة pdf الى نص :tns"=مخرجات الرساله>
</العمليه>
</نوع المنفذ>
<binding name="PdfToTextPortBinding" type="tns:PdfToText">
<operation name="pdfToTextBase64">
....................................
```

Fig. 8. Translated WSDL document

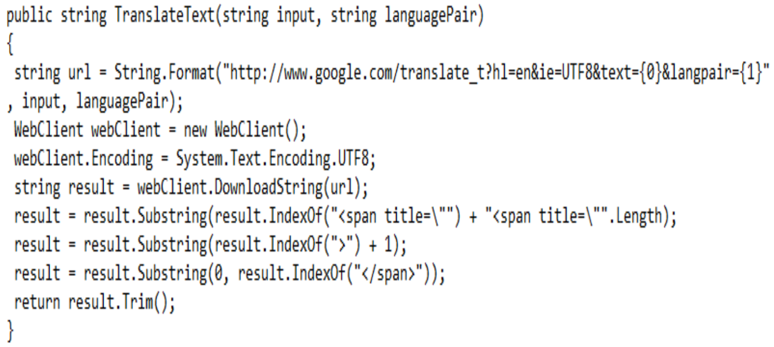
G. English To Arabic Translator Software API

As we mentioned in section 4.1.3 the retrieved WSDL documents were translated to Arabic. This process is done through a specialized software that we mentioned it in figure 6 as a black box.in this part we used a readymade software because the WSDL document contains some attributes an elements that contain a long description of the service which like:

```
<documentation>
  This service will extract the text content from a PDF file.
  It uses the pdftotext executable from Xpdf
  (http://www.foolabs.com/xpdf/).
  The text returned from this service
  often contains characters which are XML-invalid,
  therefore the text is returned in i
</documentation>
```

That needs a full translation with applying the grammars rules and machine learning techniques. For this purpose we used the Google translation APIs. Google Translate is a free statistical multilingual machine-translation service provided by Google Inc. to translate written text from one language into another. Google API client libraries are available in a number of popular programming languages like .NET and Java. The API provides access to a service and provides a single URI that acts as the service endpoint.

The following figure show a sample of the programming code using c# for accessing the Google API, the following functions has two input 1- the WSDL attribute and its value as a pair ,2- the language pair which in our case is English to Arabic. And the output is the translated text:



```
public string TranslateText(string input, string languagePair)
{
 string url = String.Format("http://www.google.com/translate_t?hl=en&ie=UTF8&text={0}&langpair={1}"
, input, languagePair);
 WebClient webClient = new WebClient();
 webClient.Encoding = System.Text.Encoding.UTF8;
 string result = webClient.DownloadString(url);
 result = result.Substring(result.IndexOf("<span title=\"") + "<span title=\"".Length);
 result = result.Substring(result.IndexOf(">") + 1);
 result = result.Substring(0, result.IndexOf("</span>"));
 return result.Trim();
}
```

Fig. 9. C# code for accessing Google API

V. Evaluation

For testing our proposed approach we searched for a registry of curated Web Services that contains a huge numbers of Web services. And we found the BioCatalogue, it is a centralized registry of curated Life Science Web Services and it has the following functions:

1) *Search (by Keywords):*
User or agent could search using any part of service name.
2) *Search (by Data):*
User or agent could search using a specific input data or output data.
3) *View Full Details of a Service:*
Also they could view the full details of a service.

The previous functions are available on the BioCatalogue.com website through a defined web interface, also there are an available API's to access all BioCatalogue.com searching, filtering, browsing data and WSDL documents. The BioCatalogue provides a set of public RESTful endpoints that allow user and agent to query the registry programmatically and integrate the data and functionality into their own scripts, workflows, apps, tools and mashups.

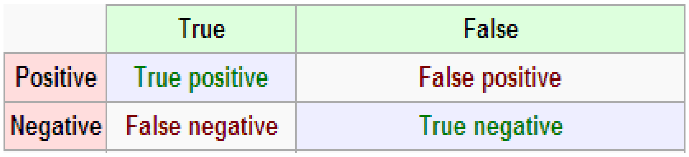
We selected a 20 English queries for a web services, we chose them based on BioCatalogue Latest Activity log, which contains the recent used web services based on users' queries. The 20 queries and their returned results were listed in the Table 1.

TABLE I. English Queries and Their results

No	English Query	No WS
1	Convert PDF to Text	13
2	Cleaning Text from unwanted classes	3
3	Text search and retrieval from large databanks	3
4	Protein Sequence Analysis on pfam	12
5	Named Entity Recognition	16

6	Biomedical Named Entity Recognizer	2
7	Document Discovery	3
8	Document Similarity	1
9	Document Clustering	1
10	Classify text	1
11	Text Mining	38
12	Sentence split	2
13	Structure Retrieval	70
14	Similarity sequence databases	1
15	Chemical text mining	1
16	Image Retrieve	5
17	Genome Information Broker	6
18	Protein sequence database query	56
19	Protein Sequence Similarity service	4
20	Protein Tertiary Structure	15

The selected English queries were translated manually to Arabic language and entered as input of our proposed model and their results were tested against the original English queries results. The Arabic queries and their correct and incorrect returned results numbers were listed in the Table 2, for each query we calculated the accuracy and precision by using the binary classification like:



	True	False
Positive	True positive	False positive
Negative	False negative	True negative

That is, the accuracy is the proportion of true results (both true positives and true negatives) in the population.

$$\text{accuracy} = \frac{\text{number of true positives} + \text{number of true negatives}}{\text{number of true positives} + \text{false positives} + \text{false negatives} + \text{true negati}}$$

On the other hand, precision or positive predictive value is defined as the proportion of the true positives against all the positive results (both true positives and false positives)

$$\text{precision} = \frac{\text{number of true positives}}{\text{number of true positives} + \text{false positive}}$$

Note: The total number of available Services for searching is **2485 services**

We supposed that the returned number of services in English language is the ideal case which represents the total number of services that query written in Arabic language should return, which exists in the third column in table 2.

Then we calculated the average accuracy and average precision for the 20Arabic queries:

AvgAccuracy = (Total accuracy/No of queries) =99.87 %

AvgPrecision = (Total precision /No of queries) =95.07%

The WSDL documents for the selected web services are translated as we mentioned before in 4.1.3 and figure 8.

And the following chart shows the similarity between the results of English query and Arabic Query:

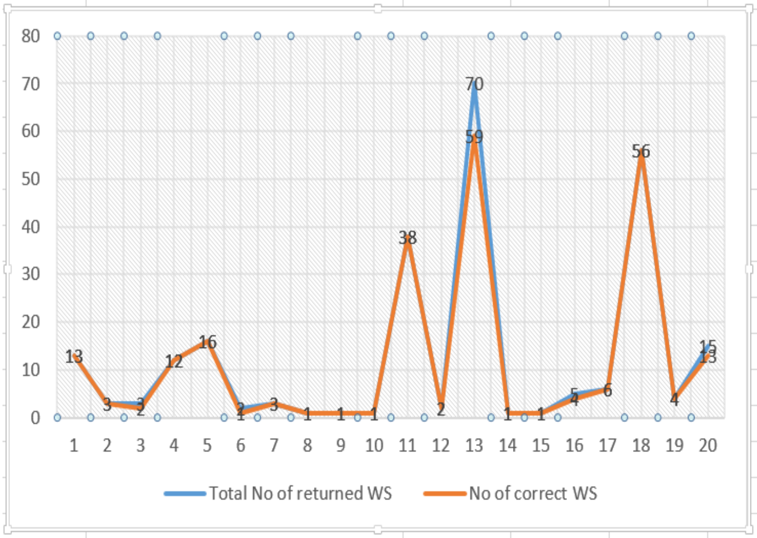


Fig. 10. English vs. Arabic Queries results

TABLE II. Arabic Queries and Their results

No	Arabic Query	Number of returned WS from English Query	Total No of returned WS	No of True Positive	No of True Negative	No of False Negative	No of False Positive	**Accuracy %**	**Precision %**
1	تحويل PDF إلى نص	13	13	13	2472	0	0	100.00	100.00
2	تنظيف النص من الطبقات غير المرغوب فيها	3	3	3	2482	0	0	100.00	100.00
3	البحث عن النص واسترجاع من قواعد البيانات الكبير	3	2	2	2482	1	0	99.96	100.00
4	تحليل تسلسل البروتين على pfam	12	12	12	2473	0	0	100.00	100.00
5	التعرف على الكيان المسمى	16	16	16	2469	0	0	100.00	100.00
6	التعرف على الطب الحيوي للكيان المسمى	2	1	1	2483	1	0	99.96	100.00
7	اكتشاف الوثيقه	3	7	2	2478	1	4	99.80	33.33
8	تشابه الوثيقه	1	1	1	2484	0	0	100.00	100.00
9	تجميع الوثيقه	1	1	1	2484	0	0	100.00	100.00
10	تصنيف النص	1	4	1	2481	3	0	99.88	100.00
11	التعدين النص	38	38	38	2447	0	0	100.00	100.00
12	انقسام الجملة	2	2	2	2483	0	0	100.00	100.00
13	هيكل استرجاع	70	59	40	2396	30	19	98.03	67.80
14	قواعد البيانات تشابه تسلسل	1	1	1	2484	0	0	100.00	100.00
15	النص الكيميائية التعدين	1	1	1	2484	0	0	100.00	100.00
16	استرجاع الصور	5	4	4	2480	1	0	99.96	100.00
17	سمسار معلومات الجينوم	6	9	6	2476	3	0	99.88	100.00
18	الاستعلام عن البروتين تسلسل قاعدة البيانات	56	56	56	2429	0	0	100.00	100.00
19	خدمه تسلسل التشابه البروتين	4	4	4	2481	0	0	100.00	100.00
20	بنية البروتين العالي	15	13	13	2470	2	0	99.92	100.00

VI. EVALUATION

Web service discovery is a very important process special after applied IR techniques which leads to IR web service discovery approach, but this limits it use to English language users only, our approach leads to expand the web service users to other languages, in our research we proposed the Arabic language approach which could be used in other languages like Indian, Chinese and so on. This may lead to further applications that could use multiple language web service and application to application different language data exchange.

Our future work to expand our approach to use the data mining techniques on the web service user's history to improve the selection mechanism.

REFERENCES

[1] A. Sajjanhar, J. Hou, Y. Zhang, Algorithm for web services matching, in: APWeb, 2004, pp. 665–670.

[2] A. Wombacher, P. Fankhauser, B. Mahleko, E. Neuhold, Matchmaking for business processes based on conjunctive finite state automata, International Journal of Business Process Integration and Management 1 (1) (2005) 3–11.

[3] C. Platzer, S. Dustdar, A vector space search engine for Web services, in: Third IEEE European Conference on Web Services, Sweden, 2005.

[4] C. Van Rijsbergen, Information Retrieval, Buttersworth, London, 1979.

[5] Chen Wu,WSDL term tokenization methods for IR-style Web services discovery,Science of Computer Programming,2011.

[6] Dina Demner-Fushman,Douglas W. Oard,The Effect of Bilingual Term List Size on Dictionary-Based Cross-Language Information Retrieval,2003.

[7] Donnla Nic Gearailt,Dictionary characteristics in cross-language information retrieval,2005.

[8] E. Christensen, F. Curbera, G. Meredith, S. Weerawarana, Web services description language (WSDL) 1.1, W3C, http://www.w3.org/TR/2001/NOTEwsdl20010315/ (March 2001).

[9] E. Christensen, F. Curbera, G. Meredith, S. Weerawarana, Web Services Description Language (WSDL) 1.1., 2001, 18/05/2007. Available: http://www.w3.org/TR/wsdl.

[10] Gina-Anne Levow , Douglas W. Oard , Philip Resnik ,Dictionary-based techniques for cross-language information retrieval,Information Processing and Management,2005.

[11] J. Zobel, A. Moffat, Inverted files for text search engines, ACM Computing Surveys 38 (2006) 1–55.

[12] Jianguo Lu, Yijun Yu,Web Service Search: Who, When, What, and How,Proceedings of the 2007 international conference,2007.

[13] Levow G-A, Oard D, Resnik P,Dictionary-based techniques for cross-language information retrieval,Information Processing & Management,2005.

[14] Mohamed Maamouri, Ann Bies, Tim Buckwalter, Wigdan Mekki,The Penn Arabic Treebank: Building a Large-Scale Annotated Arabic Corpus,2004.

[15] R. Baeza-Yates, B. Ribeiro-Neto, Modern Information Retrieval, Addison Wesley, 1999.

[16] R. Baeza-Yates, N. Fuhr, Y.S. Maarek (Eds.), Proceedings of the SIGIR Workshop on XML and Information Retrieval, 2002.

[17] Y. Wang, E. Stroulia, Flexible interface matching for web-service discovery, in: Fourth International Conference on Web Information Systems Engineering, 2003.

Actions for data warehouse success

Aziza CHAKIR
Systems Architecture Team,
Laboratory of Informatics, System
and Renewable Energy
Hassan II University - ENSEM
Casablanca, Morocco

Hicham MEDROMI
Systems Architecture Team,
Laboratory of Informatics, System
and Renewable Energy
Hassan II University - ENSEM
Casablanca, Morocco

Adil SAYOUTI
Systems Architecture Team,
Laboratory of Informatics, System
and Renewable Energy
Hassan II University - ENSEM
Casablanca, Morocco

Abstract—**Problem statement: The Data Warehouse is a database dedicated to the storage of all data used in the decision analysis, it meets the customer requirements, to ensure, in time, that a data warehouse complies with the rules of construction and manages the evolutions necessary of the information system (IS).**

Results: According to the studies carried out, we see that a system based on a data warehouse governed by the best practices of The Information Technology Infrastructure Library (ITIL) and equipped with a multi-agent system will make it possible our direction to ensure governance tending towards the optimization of the exploitation of the data warehouse.

Keywords—Information Technology Infrastructure Library (ITIL); data warehouse; governance; insufficiencies of the data warehouse; multi-agent system.

I. INTRODUCTION

The data warehouse is not merely a new practice, it is found in all the fields and corporation having data. It is a true resolution in the computing world. All the conditions are well used for a good decision making.

The raw data extracted or transformed into information are transformed by an ETL (Extraction, Transformation and Load).

The information is stored in a data warehouse to be analyzed by tools for analysis transforming this information into knowing (Figure 1).

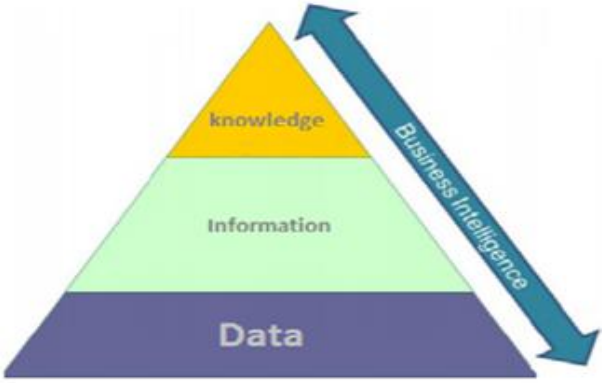


Fig. 1. Business intelligence [1]

A shallow analysis or erroneous input data may cause a wrong decision. A decisional project is prone sometimes of bugs or of dysfunction related to the bad human handling and it directly impacts the response time of a whole decisional chain.

To overcome these insufficiencies, we proposed a governance of the data warehouse based on the best practices of ITIL and of the multi-agent system.

It would be useful before presenting the method used to show the benefit of choosing ITIL and multi-agent system.

A. The benefit of choosing ITIL with data warehouse:

ITIL is a rather complete framework of reference which treats all the fields of the IT governance. . Its continuous updating and harmonization with other standards such as COBIT, ISO 27000, PMBOK and regulations such as SOX, Basel II solvency, and the possibility of using it in the data warehouses guided our choice.

B. The benefits of the governance of a data warehouse with the multi-agent system:

The governance of an information system present common point with the multi-agent system at knowing management by process [7]. This management is ideal for IT governance, which governance is not other than a set of processes in interaction between them for a better management of information technology.

II. DATA WAREHOUSE

The data used in decision making or the decisional analysis are stored to constitute a database, it is this same database which is called after data warehouse.

The tools of ETL provide the power of the data warehouse from production bases. It is only one simple copy of these data since the data warehouse transforms these last into information. Which information is transformed into knowledge through other algorithms business intelligence (Figure 2).

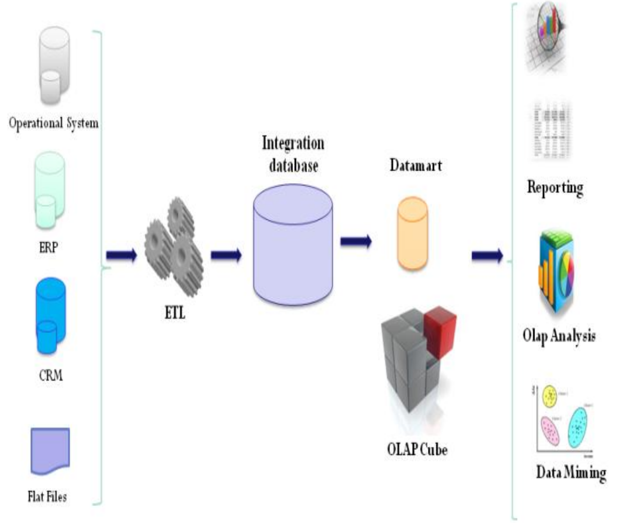


Fig. 2. The steps of feeding a data warehouse

The knowledge generated by the data warehouse is used for:

- To manage and/or predict (for example sales).
- To evaluate the risks (for example the risk customer for an insurance).
- To make a study of the behaviors of the customers in order to allow the companies to define strategies to target their customer.

A. Features of data warehouse

A data warehouse is a collection of « Subject oriented, integrated, nonvolatile, time variant collection of data in support of management decisions » [12] [13].

- Subject-Oriented:

The database is built according to the subject area, for example (customers, products, risk,...).

- Integrated:

The data comes from different production applications, can exist in all different forms.It must be integrated in order to standardize and give them a way, understandable by all users.In a data warehouse, the data must have a coding and a unique description.

- Non-volatile:

In the data warehouse, the data will not disappear and will not change with treatment over time.

- Time-Variant:

The historical data is kept in a data warehouse, The data is also stamped. We can see the evolution in time of a given value.

B. Insufficiencies of the data warehouse

The data warehouse is exposed at the risks related to [1]:

- An improper setting up, this is the case, for example, want to use this knowledge at all costs without checking the validity of data and even their volume.
- A poor quality of the data or a badly made analysis will involve erroneous results and bad decisions by the executive body of a company.
- The excellent opportunities given by a data warehouse are likely not to be exploited more, if the data warehouse set up causes changes for the users or of the bugs during their use.
- The need to change the decision in the case of development of a company (for example creation of new services).

It is clear that a data warehouse is not easy to implement and to maintain in operational condition.

III. Governance

The governance of information systems [2] [3] [11], is the procedure that defines the way organizations are able to align IT strategy with business strategy, and to ensure that companies remain on track to achieve their goals, and implement good ways to measure performance.

IT governance provides effective, efficient and compliant computer to enable an organization to achieve its objectives use.

A data-processing framework of governance answers some key questions, such as the way in which the computer department functions as a whole and that which the management of the key indicators needs.

A. ITIL is a normative reference frame

ITIL (Information Technology Infrastructure Library) [4] [5] is a framework of best practices for the delivery of IT services. It helps to improve efficiency and reduce risk.

ITIL provides a methodological approach consisting of a series of modules to help companies and organizations to improve the use of IT resources.

ITIL consists of five modules, all modules will manage an IT service and align IT services with objectives of a company. The five modules are:

- Strategy of the services.
- Design of the services.
- Transition from the services.
- Exploitation of the services.
- Continuous improvement of the services.

B. ITIL and IT service support

Support of IT services using ITIL all aspects to ensure that the field of information technology can support the IT applications that provide business functions and can guarantee the continuity, availability and quality of service to users. [15]

Support service defined by ITIL is provided through five key processes and function and are used as following:

- Service Desk (function)
- Configuration Management (process)
- Change Management (process)
- Release Management (process)
- Incident Management (process)
- Problem Management (process)

The diagram below summarizes the key aspects of the methodology support service defined by ITIL (Figure 3).

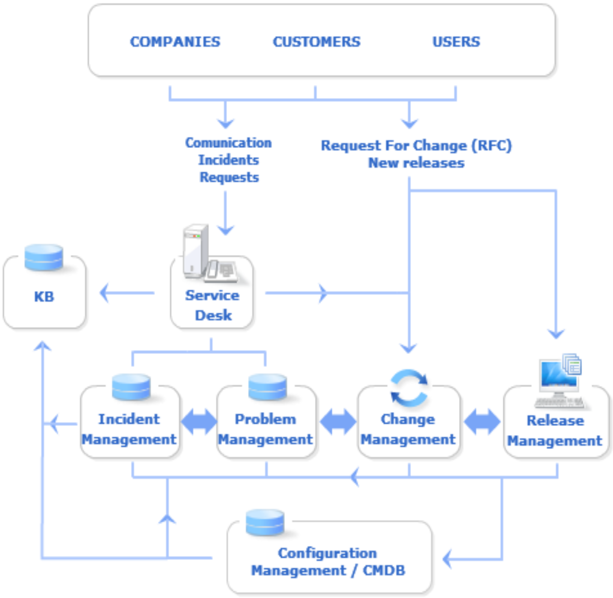


Fig. 3. The methodology of support of service defined by ITIL [1]

C. *ITIL and IT service delivery*

In order to monitor and improve the quality of the applications of the information system on levels of customer service, ITIL offers the following five processes:

- Management of service levels
- Capacity Management
- Availability Management
- IT service continuity management
- Financial management service

D. *The role of ITIL in data warehouse*

To keep the data warehouse in good condition, we must make good needs analysis, and conceive well the data which will be to use by it, and maintain the data warehouse provide a high quality service to customers. And ensure good management problems encountered, the infrastructure management, management of information dissemination, performance tuning database and the level of service provided.

ITIL is the best practice one to implement a data warehouse to work properly [6][8][9][10].

The following table (Table 1) shows the usefulness of ITIL processes to keep the data warehouse in good condition.

TABLE I. Role of ITIL in data warehouse

ITIL processes	Role of ITIL in data warehouse
Service Centre	Process to ensure the processing of all user expectations that these are simple requests or malfunctions caused by the data warehouse.
Incident management	The aim of incident management is to restore data warehouse in the shortest time, with minimal impact on users.
Problem Management	Process to optimize the level of service by analyzing the real causes of malfunctions and there by anticipating corrective action to address the shortcomings in organizing and controlling the use of resources.
Change management	Process describing the activities to quickly and efficiently conduct all changes to minimize the risk of negative impact of these changes on the quality of service.
Management put into production	Process to coordinate all activities related to the storage, management, distribution and implementation of data warehouse
Availability management	This process ensures a level of availability of the data warehouse to customers in accordance with the contract services and remaining financially viable [6].
IT service continuity management	This process ensures the continuity of the data warehouse in the event of incident [6].
Service level management	This process maintains the planning, contracting, implementation and monitoring of services and service levels, working with clients responsible for this activity and providers responsible for providing the service [6].

IV. MULTI-AGENT SYSTEM

A multi-agent system is a distributed software system consisting of several autonomous entities with different interests - agents, occurring at the same time, sharing common resources and communicating them.

Multi-agent systems can reduce the complexity of solving a problem by dividing the required subsets namely, combining independent to each of these subsets intelligent agent and coordinating the activity of these agents.

Modeling of the proposed solution is based on the principle of Multi-agent systems is: "Everyone must cooperate to achieve the same goal."

The proposed architecture is composed of the following agents:

- User Agent

Cognitive agent can communicate, intervene and monitor service-center Agent and process agent.

- Service-Center Agent

Reactive agent that reacts when the action is required.

- Process Agent

Hybrid agent, by stimulating service-center agent and cooperation with the user agent and the business agent, it takes a decision.

- Business Agent

Reactive agent, depending on the situation, it interacts with the process agent.

- Knowledge Base Agent

Reactive agent whose task is the retrieval of information from the knowledge base.

A. *Functional specifications*

A service is triggered by an event which is detected by the sensor of a process.

When the process started, it draws data from the knowledge base for possibly trigger other processes and generates an audit report.

Monitoring services will be measured in% for carrying out checks of all active processes.

The historical services provide a basis for an intelligent system.

The following diagram shows the operation of a process (for example "change indicator") (figure 4):

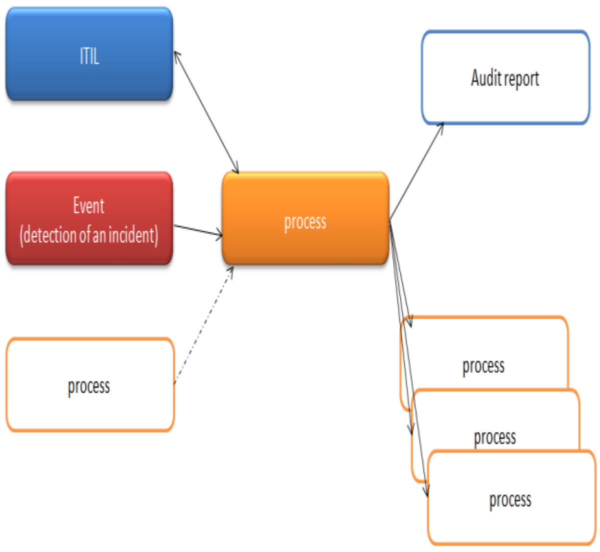


Fig. 4. Example-change indicator

V. CONCLUSION

According to the study there are gains to be made through better governance of a data warehouse. Using our method, we ensure customers satisfaction, quality service, and maintenance of a data warehouse in good condition.

Using the best practices of ITIL by professionals allows to highlight the operations that lead to improvements that can recognize weaknesses in our control.

Hence the need for a decision-making system, data warehouse, most reliable way to get to identify its problems seen to reduce the disturbances due to a malfunction in an IT services company, and this can only be done by governance of decision support system, data warehouse.

The proposed method suggests tactics for each blocking position to maintain the operability of a data warehouse.

We expect the development of a generic platform to represent the proposed method.

References

[1] Maxime Poletto, " L'informatique décisionnelle-Thèse Professionnelle ", 01 juin 2012.

[2] Morley+Al, " Processus métiers et S.I. - Gouvernance, management, modélisation - 3e édition".

[3] Jamal Skiti et Hicham Medroumi, "La Gouverance des Technologies de l'Information à base du Système Multiagent et le référentiel COBIT".

[4] Pascal Delbrayelle, http://www.itilfrance.com.

[5] Nicolas Dewaele, "L'ITIL: Un référentiel pour la qualité des systèmes d'information", 23 mars 2011.

[6] Tariq Rahim Soomro et Hasan Yousef Wahba, "Role of Information Technology Infrastructure Library in Data Warehouses", 2011.

[7] Jamal Skiti et Hicham Medroumi, "Nouvelle Méthodologie de la Gouverance des Technologies de l'information à base du Système Multi-agent".

[8] Yves B.Desfossés , Claude Y.Laporte , Alain April et Nabil Berrhouma, " Méthode d'amélioration des services de TI, basée sur ITIL, dans les entreprises québécoises ", septembre 2008.

[9] Aziza Chakir, Hicham Medromi et Adil Sayouti, "La gouvernance du système d'information à base des bonnes pratiques d'ITIL V3", novembre 2012.

[10] STEIGMEIER Alexandre, "Comment articuler les différentes normes et méthodes".

[11] Bruno Claudepierre, "Conceptualisation de la Gouvernance des Systèmes d'Information, Structure et Démarche pour la Construction des Systèmes d'"Information de Gouvernance", 10 décembre 2010.

[12] W. H. Inmon, "Building the Data Warehouse",4 edition (October 7, 2005).

[13] Elzbieta Malinowski, Esteban Zimanyi, "Advanced Data Warehouse Design: From Conventional to Spatial and Temporal Applications".

AUTHORS

Aziza Chakir is an engineer in computer science from the ENSIAS, Mohammed V – Souissi University in July 2009, Rabat, Morocco. She prepare her thesis at ENSEM, Hassan II University, Casablanca, Morocco.

Hicham Medromi received the PhD in engineering science from the Sophia Antipolis University in 1996, Nice, France. Heis responsible of the system architecture team of the ENSEM Hassan II University, Casablanca, Morocco. His actual main research interest concern Control Architecture, Architecture of System and Software Architecture Based on Multi Agents Systems and Distributed Systems. Since 2003 he is a full professor for Control systems and computer sciences at the ENSEM, Hassan II University, Casablanca. He managed eight Research projects and he has published several Patents and publications in international journals and conferences.

Adil Sayouti received the PhD in computer science from the ENSEM, Hassan II University in July 2009, Casablanca, Morocco. In the same year he received the price of excellence of the best sustained thesis in 2009. In 2003 he obtained the Microsoft Certified Systems Engineer (MCSE). In 2005 he joined the system architecture team of the ENSEM, Casablanca, Morocco. His actual main research interests concern Remote Control over Internet Based on Multi agents Systems.

Intelligent Ambulance Traffic Assistance System

RONOJOY GHOSH
IT, Institute of Engineering and Management

VIVEK SHAH
ECE, Institute of Engineering and Management

HITESH AGARWAL
CSE, Institute of Engineering and Management

ASHUTOSH BHUSHAN
ECE, Institute of Engineering and Management

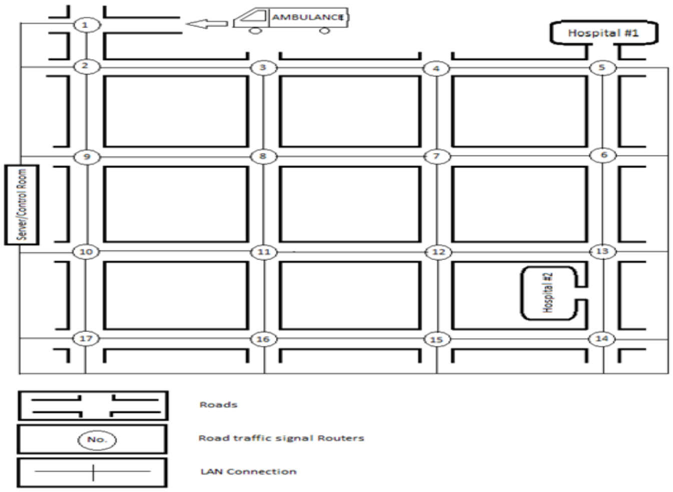
PRASUN KANTI GHOSH
CSE, Institute of Engineering and Management

Abstract—**With the increase in traffic road density, several causalities occur due to delay in taking a patient to the hospital in an ambulance. In this paper, we have developed an algorithm to find the shortest path to reach the required destination. As required the software will identify the present location of the vehicle and ask the user for the destination. Then it will show all the available paths, highlighting the shortest one or in several cases the most optimum one. Further we made the traffic signals automated for special vehicles like an ambulance or a fire-engine such that the signals will go green for the ambulance as it comes in the vicinity of the traffic signal, thus providing them with a clear path to reach its destination. The original signal is restored as soon as the ambulance goes undetected by the Bluetooth scanner of the traffic signal.**

***Keywords*—python; ambulance; wireless; Bluetooth; cryptography;**

MOTIVATION

In India rapid growth of population coupled with high rate of industrialization has resulted in unmanageable increase in traffic volume, especially in metropolitan cities and urban areas. Due to this increase in traffic density several valuable lives are lost due to delay in receiving medical attention. So we designed a system which prioritizes emergency vehicles like ambulances, fire engines and provide them with a congestion free path to reach its destination as soon as possible.



PROBLEMS

One of the most challenging problems of urban civilization is directly or indirectly related to population explosion. Traffic congestion being one of the most persistent one. Not only it wastes our valuable time but also in some cases situations can go critical.

- Traffic congestion hampers the speed of vehicles which also include emergency vehicles like ambulances, police van and fire engines whose delay can put life of many at risk.
- Unnecessary waiting at the traffic signals due to unequal traffic density
- Absence of knowledge about the routes of a city.
- Handling huge traffics can get daunting at times.

It's the emergency services which pay the maximum price when caught in traffic jams especially the ambulance services where situations can be very critical. To solve this problem we have come up with the solution of Intelligent Ambulance Traffic Assistance system using secure wireless networks.

Using this technology we make traffic signals automatically green as any ambulance comes in its vicinity, thereby minimizing the unnecessary time spent in traffic jams. As a result it gets a clear path to carry the patient to a nearby hospital which can be at times quite vital to save one's life. Moreover we find all the available paths from the current position of the vehicle to the hospital, highlighting the shortest or optimal path., this facility can also make up for the meagre knowledge of routes of the driver. It also solves the problem of language barrier which is experienced by many drivers who are new to a completely cultural diverse location where our program proves to be quite useful.

Since the changing of signals is completely automated we make the task of traffic operators quite hassle free and comfortable.

This facility further enables us to deal traffic according to its density. It's unnecessary to keep a whole lot of vehicles waiting at a 'red' signal for 60secs than letting few vehicles pass through 'green' signal for that equal amount of time. The path which has higher traffic density faces 'red' signal for lesser time and more of green signal and vice versa. This greatly enhances the mobility of the vehicles

As automated systems are taking over the manual ones due to their increased efficiency and human error free nature, introduction of this technology takes it to an advanced level.

I. Introduction

We aim to develop an automatic traffic control system which can function independent of any outside help. Our traffic control system is equipped with a distinct feature especially for ambulance through which it is being assigned priority in terms of getting a green signal ahead of normal vehicles. Along with that we are also determining the optimum path between any source and destination. But the most crucial point is that we are providing all these facilities with the help of very cheap and widely available technologies, which makes us distinct from others.

We implemented the following technologies in this project:

- Python (as coding language)
- Bluetooth(as a mode of communication)
- Cryptography(to enhance system security)

What is Python?

- Python is a general-purpose, high-level programming language whose design philosophy emphasizes code readability. Python language stands out in comparison with respect to other programing languages as it:
- Is compact.
- Can be packaged into standalone executable file which can cater to our several needs by using third party tools
- Provides a simpler and better way to represent data in graphical form-
- Modules like matplotlib, visual provides us with the facility of mapping data in 2-D and 3-D form.
- Is an interpreted language, that allows for rapid, flexible, exploratory software development

Why python?

Python language stands out in comparison with respect to other programing languages in terms of its vast array of standard library and code readability. We chose it over other languages as it is more compact and also by using third party tools python code can be packaged into standalone executable file which can cater to our several needs. We have used third party modules like Bluetooth, matplotlib, pysqlite, cipher and visual. Bluetooth module helps us to perform the scanning and tracking operation. Python provides a simpler and better way to represent data in graphical form, in comparison to any other programing language. We did the implementation of the above with the help of module matplotlib. We have connected database to python with the module pysqlite as we need database connection in our project to bring coordination between the traffic signals. To provide security to the files used in the project from unscrupulous elements, we used the module cipher. By this module, we can easily form a sphere or other known figure in only a single statement, which in comparison, takes several steps in other programming language to accomplish.

Python over Java:

- Concise Coding style: The code in Python is typically much more concise than that of Java, with much lesser verbosity.
- Dynamic Typing: No requirement of declaring data types in Python making sure that the inheritance hierarchies especially for all the interfaces and implementations are well laid out.
- Built in language capabilities: Python has more built in language capabilities than Java. Items such as list comprehensions, ability to deal with functions as first class objects gives a broader vocabulary to work with.

What is Bluetooth?

- Bluetooth is a wireless technology standard for exchanging data over a short distance in a very efficient and lucid way. It is a technology standard for exchanging data over short distances (using short-wavelength radio transmissions in the ISM band from 2400–2480 MHz) from fixed and mobile devices, creating personal area network (PANs) with high levels of security. Created by telecom vendor Ericson in 1994 it was originally conceived as a wireless alternative to RS-232 data cables.

Why Bluetooth?

- It is very cheap and easily available- A Bluetooth adapter costs very less (around 200 INR) and can be found easily in the market. Therefore its maintenance can be done easily.
- It can connect several devices, overcoming problems of synchronization.
- It is very easy to install- Unlike its other connecting devices it consumes very less power so it can operate without an external source.
- Limited area of access proves to be quite an advantage for our project.
- Does not interfere with normal signalling of devices.
- Since our project mainly deals with ambulance and other emergency vehicles which can contain sophisticated and sensitive devices .These devices can get easily affected by the interference of an external signal and Bluetooth just avoids that.

Why better than Wi-Fi?

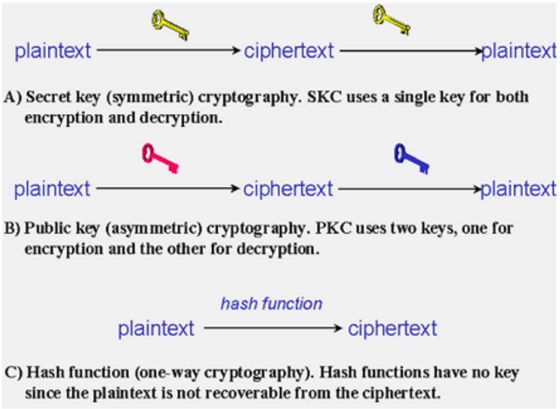
- Cost effective-
- Cost of setting a Bluetooth in comparison to Wi-Fi is much less.

- Interferes with external signalling of devices-
- At the time of functioning, Wi-Fi devices tend to interfere with signalling of surrounding devices which can prove quite harmful in several cases when dealing with certain sophisticated and life saving devices.
- Lower power consumption and easy to maintain-
- A portable device's primary need is that it should consume less amount of power, but Wi-Fi devices require larger amount of power to function properly which makes it quite inappropriate within the context of our project.

What is cryptography?

Cryptography is the science of information security. It is about construction or formation of some basic protocols to overcome the influence of adversaries. Modern cryptography concerns itself with the following four objectives:

- Confidentiality -the information cannot be understood by anyone for whom it was unintended.
- Integrity -the information cannot be altered in storage or transit between sender and intended receiver without the alteration being detected.
- Non-repudiation -the sender of the information cannot deny at a later stage his or her intentions in the creation or transmission of the information.
- Authentication -the sender and receiver can confirm each other's identity and the origin & destination of the information.



A) Secret key (symmetric) cryptography. SKC uses a single key for both encryption and decryption.

B) Public key (asymmetric) cryptography. PKC uses two keys, one for encryption and the other for decryption.

C) Hash function (one-way cryptography). Hash functions have no key since the plaintext is not recoverable from the ciphertext.

Fig. 1. Encryption and Decryption in Cryptography

What is the need of cryptography?

Our project deals with full automation of traffic signals which is vulnerable to any kind of anti-social activities which can prove quite troublesome. At the time when database is edited through scanner which is the deciding factor in management of the traffic lights any interference from any foreign unscrupulous element will make the whole situation more chaotic and can disrupt the whole traffic system. So by securing it we are trying to avert or avoid any such activities.

II. Related Work

In this section we mention about previous works which have motivated us in implementing this project. We have seen a number of projects which used Bluetooth Scanning devices for various purposes. Coding of various projects is also done using python. The combination of Bluetooth and python has helped in developing different types of projects. In the paper 'WRife:a' wireless epidemic data collection protocol suitable for medical monitoring' by students of Texas University, we see that patient's health related data is collected using medical sensors. Fixed and mobile radio devices are used for disseminating information from medical sensors to the servers. In this project Wi Fi could have been used but RF radiation emitted by it interferes with medical devices and may also cause health hazards. The bandwidth requirement of Bluetooth is much less than that of Wi fi and is also low power consuming device, thereby reducing interference and safety concerns. Moreover in this project poisoning where attackers inject corrupted data causing the loss or modification of original message sample, various cryptographic ciphers have been used.so we have also encrypted our database file before sending it to the client. The client can decrypt the file using the key. In another project 'A deadline driven epidemic data collection protocol suitable for tracking inter personnel rendezvous' we can see the application of Bluetooth for peer to peer wireless data collection algorithm.

In our project we have three components-ambulance which acts as the client, scanners which act as Bluetooth access points and master servers. When the ambulance sends signal to the scanner in order to get the database for finding the shortest path leading to the hospital, the scanner sends the database to the ambulance in an encrypted format. This similar approach is observed in another work 'In-Building Location using Bluetooth'. In this project the location of any mobile device can be detected using Bluetooth scanners. The received signal strength from each coordinate is sent to the server by the scanners. The server has a map of RSSI (Received Signal Strength Indication) at different coordinates. Thus it gives the deduced location of the mobile device by the use of the received RSSI and triangulation technique.

III. Methodology

A. Administration

1) *Flow Diagram:*

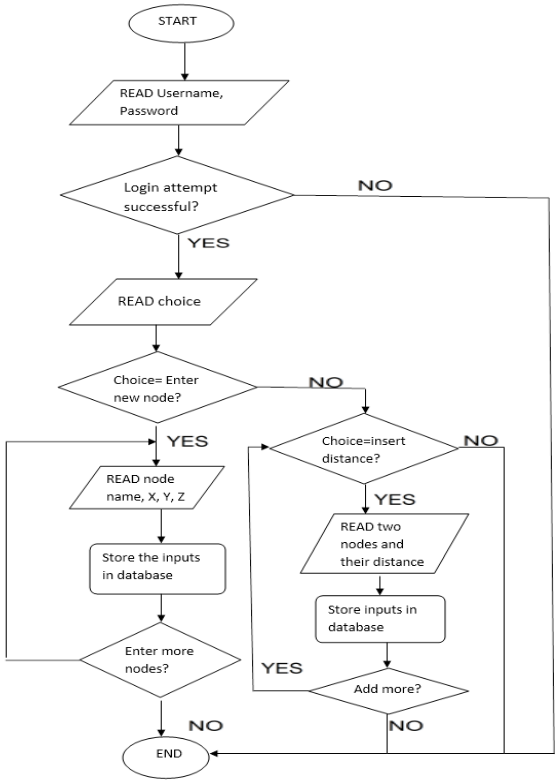


Fig. 2. Flow Diagram for Administration

2) Step Algorithm

Step 1: START

Step 2: Login window opens

Enter user name: name

Enter password: pass

Step 3: If username exist in table:

If password==pass:

//login successful

//Another window opens

To enter new node GOTO Step 4

To insert distance between two nodes GOTO Step 7

Else:

//login unsuccessful

GOTO Step 2

//ENTER A NEW NODE

Step 4: click**”** ENTER NEW NODE” button in new window

//another window opens

GOTO Step 5

Step 5: Enter new node: node

Enter X co-ordinate: x

Enter Y co-ordinate: y

Enter Z co-ordinate: z

“Submit” button is pressed

Step 6:”ENTER ANOTHER NODE” button is pressed

//to insert another node

//another window opens

GOTO Step 5

//INSERT DISTANCE BETWEEN NODES

Step 7:”enter distance between paths” button is pressed

//another window opens

GOTO Step 8

Step 8: Enter source: source

Enter destination: destination

Enter distance: distance

“Submit” button is pressed

Step 9:”ENTER ANOTHER DISTANCE” is pressed

//to insert distance of other two nodes

GOTO Step 8

B. Ambulance

1) Flow Diagram

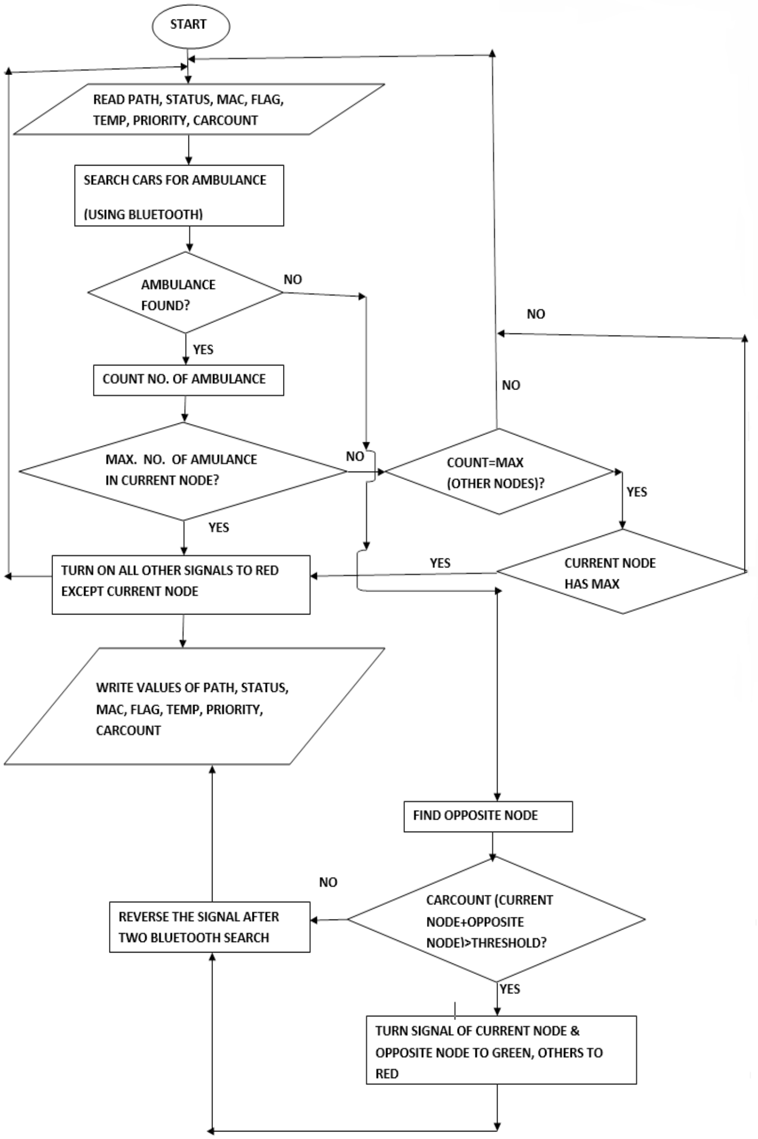

START
READ PATH, STATUS, MAC, FLAG, TEMP, PRIORITY, CARCOUNT
SEARCH CARS FOR AMBULANCE (USING BLUETOOTH)
AMBULANCE FOUND?
NO
YES
COUNT NO. OF AMBULANCE
NO
NO
MAX. NO. OF AMULANCE IN CURRENT NODE?
NO
COUNT=MAX (OTHER NODES)?
YES
YES
YES
TURN ON ALL OTHER SIGNALS TO RED EXCEPT CURRENT NODE
CURRENT NODE HAS MAX
WRITE VALUES OF PATH, STATUS, MAC, FLAG, TEMP, PRIORITY, CARCOUNT
FIND OPPOSITE NODE
NO
REVERSE THE SIGNAL AFTER TWO BLUETOOTH SEARCH
CARCOUNT (CURRENT NODE+OPPOSITE NODE)>THRESHOLD?
YES
TURN SIGNAL OF CURRENT NODE & OPPOSITE NODE TO GREEN, OTHERS TO RED

Fig. 3. Flow Diagram for Ambulance

2) Step Algorithm

Step 1: START

Step 2: read path, status, mac, flag, temp, priority, car count from file

Step 3: search cars (Bluetooth devices) at a particular node

Step 4: if mac address of ambulance is found in search:

GOTO Step 5

Else:

GOTO Step 13

Step 5: Count the number of ambulance present in current node

Step 6: find number of ambulance present in every other nodes

Step 7: if maximum number of ambulance is present in current node:

GOTO Step 8

Else:

GOTO Step 9

Step 8: Turn all node-signal RED except current node, which is turned GREEN

GOTO Step 12

Step9: if number of ambulance present in current node is same as any other node and that is maximum value of ambulance in any node:

GOTO Step 10

Else:

GOTO Step 2

Step 10: check priority of each node

Step 11: if priority of current node is maximum:

GOTO Step 8

Else:

GOTO Step 2

Step 12: write updated values of path, status, mac, flag, temp, carcount into file

GOTO Step 2

Step 13: find opposite node of current node

Step 14: check if sum of carcount of current node and opposite node is greater than threshold

Value then:

GOTO Step 15

Else:

GOTO Step 16

Step 15: turn on current node and opposite node-signal to GREEN and others to RED

GOTO Step 12

Step 16: reverse node-signal of each node after 2 bluetooth search time (1 bluetooth search time

requires approximately 8 seconds)

GOTO Step 12

C. Procedure Path

1) Flow Diagram

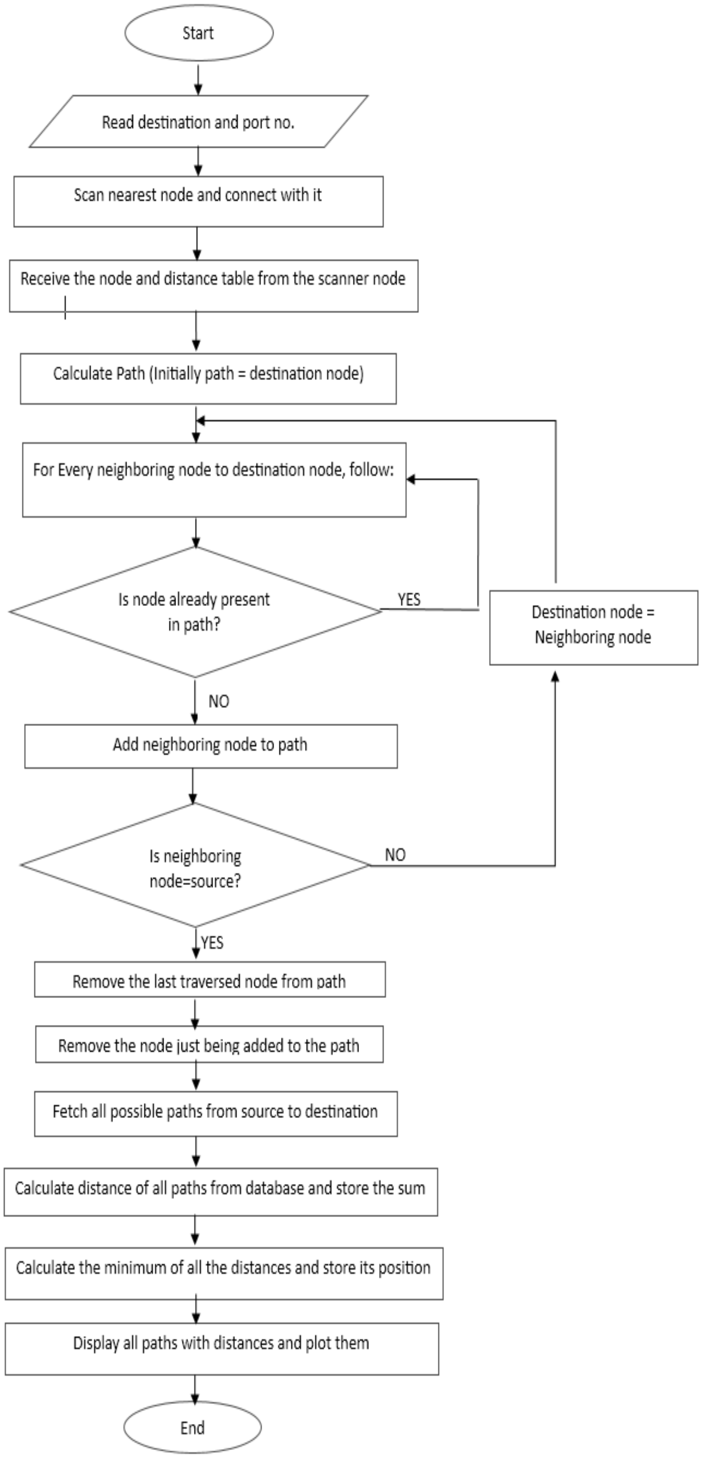

Start
Read destination and port no.
Scan nearest node and connect with it
Receive the node and distance table from the scanner node
Calculate Path (Initially path = destination node)
For Every neighboring node to destination node, follow:
Is node already present in path?
YES
Destination node = Neighboring node
NO
Add neighboring node to path
Is neighboring node=source?
NO
YES
Remove the last traversed node from path
Remove the node just being added to the path
Fetch all possible paths from source to destination
Calculate distance of all paths from database and store the sum
Calculate the minimum of all the distances and store its position
Display all paths with distances and plot them
End

Fig. 4. Flow Diagram for Procedure Path

2) Step Algorithm

Step 1: Start.

Step 2: Read the destination and the port number through which you want to communicate.

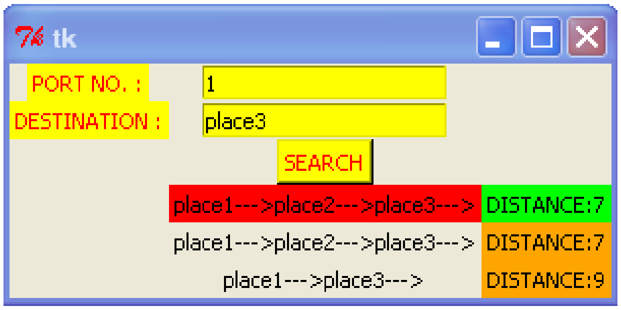
Step 3: Scan the nearest node/traffic signal scanner.

Step 4: Connect to the node scanner with the

chosen port number, and send it a message.

Step 5: Receive the node and distance tables (database) in encrypted form from the scanner node, along with the name of the scanner.
Step 6: Decrypt the tables to get the actual values.
Step 7: Calculate the path using the procedure Calculate Path (source node name, destination node name and path). Initially variable path contains the destination node.
//PROCEDURE CALCULATE PATH
Step 8: For every neighbouring node to the current destination node, carry out the following steps.
Step 9: If this node is already present in the path:
Go to step 8 and continue execution with the next value.
Else:
Go to step 10.
Step 10: Add the neighbouring node to the path.
Step 11: If this neighbouring node is the source:
Go to step 12.
Else:
Go to step 15.
Step 12: Store the path.
Step 13: From the path, remove the last node traversed.
Step 14: Go to step 8.
Step 15: Go to procedure CALCULATE PATH with only the destination node parameter being replaced by the neighbouring node.
Step 16: Remove the node just being added to the path.
Step 17: Return from the procedure CALCULATE PATH, fetching all the paths possible to go from the source to the destination.
Step 18: For all paths obtained between the source and the destination through the above procedure, calculate their respective distances with the help of the database.
Step 19: Store these distances in a list for later usage.
Step 20: Calculate the minimum distance of all the distance of the paths and store its position.
Step 21: Display all the paths, along with its distances in the window, where the paths are searched.
Step 22: Using the modules matplotlib (pyplot) and visual, plot the two dimensional and three dimensional views of the paths with respect to the co-ordinates of the nodes obtained from the database. The minimum distance path is differentiated by colouring it red while the rest are blue.
Step 23: End.

IV. Results



PORT NO. : 1
DESTINATION : place3
SEARCH

```
place1--->place2--->place3--->  DISTANCE:7
place1--->place2--->place3--->  DISTANCE:7
place1--->place3--->            DISTANCE:9
```

Fig. 5. Every possible path with distance along with the shortest path separately.

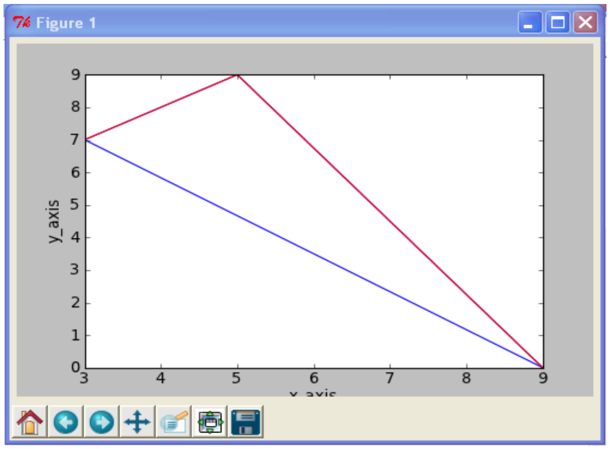


Fig. 6. Every paths along with shortest path in 2D plot.

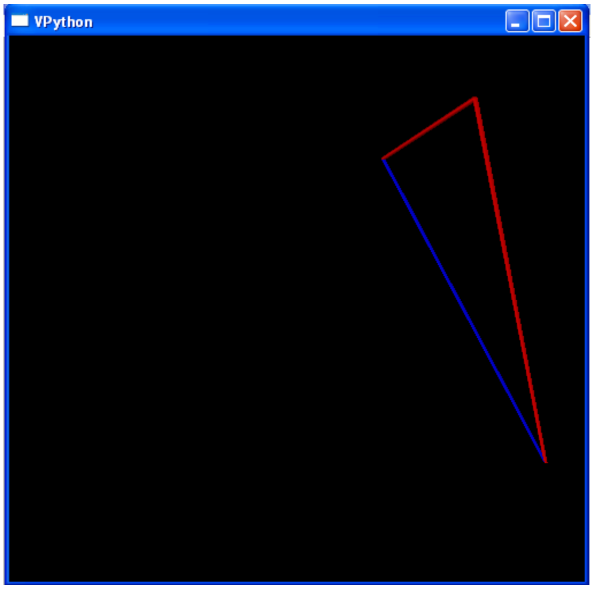


Fig. 7. Every path along with shortest path in 3D plot

V. Future Work

Our project can be enhanced by working in the following area,

- Encryption
- Tracking of vehicles
- Generalizing the ambulance concept
- Stoppage time

1) Encryption: We can increase the level of security of our system by making the encryption key even more secure because only the secrecy of the key provides security and it's better to assume that the intruder knows the system. We can make the system even more secure by encoding the encryption key. We can do it by using a separate key (second level key) to encrypt the basic encryption key. Hence to *decrypt the file containing the information we have first decrypt the basic key using second level key and then decrypt the file using the decrypted basic key. This is known as second level encryption-decryption. Similarly we can increase the level of encryption-decryption by increasing the number of keys, each key, barring the final key, is encrypted.*

2) If we use AES instead of DES it will enhance the level of security of the system. We can also increase the size of the key from 128 to 256 bits but it will require higher hardware configuration which in turn increases the cost of the project.

3) Tracking of vehicles: we can track any vehicle as every vehicle will have an inbuilt Bluetooth device having unique MAC id. The Bluetooth scanners installed in the traffic signals will monitor the vehicle and the vehicle's path can be derived by fetching the time and the position of the Bluetooth scanners which are scanning the vehicle. Hence, we can derive the path of the vehicle by comparing the scanning time of the different Bluetooth scanners.

4) Generalizing the ambulance concept: In our project the system give special preference to ambulance, where the signal allows its smooth mobility, blocking other signals. This concept can also be applied to fire engines and for the vehicles of VVIP's. The work load of traffic police is substantially reduced, thereby, less number of traffic police can be deployed.

5) Stoppage time: We often encounter road blocks during maintenance, special occasions, procession etc. To avoid this kind of trouble we can introduce a new concept of "stoppage time", which is the time difference between the actual time and the ideal time taken by a vehicle to cover a certain path. We can incorporate this stoppage time concept along with the shortage path algorithm to find the real time optimal path.

VI. Conclusion

From a proper analysis of positive points and constraints of the system it is inferred that the system is working as per the objectives of the project. Installation and maintenance of the system is cost effective and takes less time. The system-user interface is user friendly and does not require specialized training or skills to operate it.

The project has been designed to substantially enhance the performance by ensuring smooth mobility of emergency services (like ambulance, fire engines, etc.).The implementation of the algorithm is done in such a way that it not only paves way to emergency vehicles but it's auto reinstatement of the older status of traffic light helps in smooth transition of traffic along the road. The system also reduces the workload of traffic personnel as it totally automates the whole prospect of traffic signalling which also greatly reduces the domain of error. We have also equipped it with an algorithm which provides the user with the shortest possible path between destination and source which is the biggest asset in this era where people consider time as money.

Being an automated signalling system it eliminates the chances of human error which often results in road accidents and mishaps.

As discussed earlier, this project transforms the shortcomings (in terms of range and scanning time) of Bluetooth Technology into its strength thereby consolidating its applicability as the time lag between detection of two vehicles has to be wide enough to avoid any complications. A scan time of usually 8 seconds also provides us with adequate time for reinstating of older status of traffic lights.

Thus this project is practically feasible, economically viable, and reliable in nature. It's robust as well as easy to handle mechanism makes it easy and quite simple to be understood and brought in use by the masses. Summing up we can say that this project with its ready to apply technology and cheap installation charges invariably finds its application in our traffic signalling system.

An improvisation of the project and subsequent modification of the system can serve our purpose as and when needed in near future.

VII. Acknowledgement

The satisfaction that accompanies that the successful completion of any task would be incomplete without the mention of people whose ceaseless cooperation made it possible, whose constant guidance and encouragement crowned all efforts with success.

We are grateful to our project guide Mr. Avranil Tah and Mr. Abhishek Mallik for the guidance, inspiration and constructive suggestion that helped us in the preparation of this project.

We would like to thank E-School Learning and Institute of Engineering and Management for allowing us to do this project successfully.

References

[1] Avranil Tah, "A deadline-driven epidemic data collection protocol suitable for tracking interpersonal rendezvous" (January 1, 2010). ETD Collection for University of Texas, El Paso. Paper AAI1483985.
http://digitalcommons.utep.edu/dissertations/AAI1483985

[2] Cryptography and Data Security (1982)by Dorothy E. Denning , Peter , J. Denning

[3] http://www.edmunds.com/car-technology/what-the-heck-is-bluetooth-and-why-should-i-care.html
Motivates us why we should use Bluetooth as our main area of work especially in automobile field

[4] Python Geospatial Development
By Erik Westra ISBN 13: 978-1-84951-154-4 Packet publishing, 508 pages (December 2010)
Build a complete and sophisticated mapping application from scratch using Python tools for GIS development.

[5] Learning Python, 5th Edition Powerful Object-Oriented Programming By Mark Lutz Publisher: O'Reilly Media Pages: 1600 Explore Python's major built-in object types such as numbers, lists, and dictionaries Create and process objects with Python statements, and learn Python's general syntax model

[6] NumPy Beginner's Guide - Second Edition Language: English Paperback: 234 pages [235mm x 191mm] Release Date: November 2011 ISBN 13: 9781849515306 Author(s): Ivan Idris Topics and

Technologies: All Books, Big Data and Business Intelligence, Data, Beginner's Guides, Open Source, Python

[7] ABA journal-talking Bluetooth Posted Jan 2, 2004 4:27 AM CDT By David Beckman and David Hirsch http://www.abajournal.com/magazine/article/talking_bluetooth/

[8] Applied Cryptography Second Edition Bruce Schneier John Wiley & Sons, 1996 ISBN 0-471-11709-9

[9] Head First Python by Paul Barry Python syntax Setting up your environment Sharing code with PyPi Data manipulation File handling

Investigate the Performance of Document Clustering Approach Based on Association Rules Mining

Noha Negm
Math. And Computer Science Dept.
Faculty of Science, Menoufia University
Shebin El-Kom, EGYPT

Passent Elkafrawy
Math. And Computer Science Dept.
Faculty of Science, Menoufia University
Shebin El-Kom, EGYPT

Mohamed Amin
Math. And Computer Science Dept.
Faculty of Science, Menoufia University
Shebin El-Kom, EGYPT

Abdel Badeeh M. Salem
Computer Science Dept. Faculty of Computers and
Information, Ain Shams University
Cairo, EGYPT

Abstract—**The challenges of the standard clustering methods and the weaknesses of Apriori algorithm in frequent termset clustering formulate the goal of our research. Based on Association Rules mining, an efficient approach for Web Document Clustering (ARWDC) has been devised. An efficient Multi-Tire Hashing Frequent Termsets algorithm (MTHFT) has been used to improve the efficiency of mining association rules by targeting improvement in mining of frequent termset. Then, the documents are initially partitioned based on association rules. Since a document usually contains more than one frequent termset, the same document may appear in multiple initial partitions, i.e., initial partitions are overlapping. After making partitions disjoint, the documents are grouped within the partition using descriptive keywords, the resultant clusters are obtained effectively. In this paper, we have presented an extensive analysis of the ARWDC approach for different sizes of *Reuter's* datasets. Furthermore the performance of our approach is evaluated with the help of evaluation measures such as, *Precision*, *Recall* and *F-measure* compared to the existing clustering algorithms like Bisecting K-means and FIHC. The experimental results show that the efficiency, scalability and accuracy of the ARWDC approach has been improved significantly for *Reuters* datasets.**

Keywords—Web Document Clustering; Knowledge Discovery; Association Rules Mining; Frequent termsets; Apriori algorithm; Text Documents; Text Mining; Data Mining

I. INTRODUCTION

The internet has become the largest data repository, facing the problem of information overload. The existence of an abundance of information, in combination with the dynamic and heterogeneous nature of the Web, makes information retrieval a tedious process for the average user. Search engines, Meta-Search engines and Web Directories have been developed in order to help the users quickly and easily satisfy their information need. The Search engine performs exact matching between the query terms and the keywords that characterize each web page and presents the results to the user. These results are long lists of URLs, which are very hard to search. Furthermore, users without domain expertise are not familiar with the appropriate terminology thus not submitting the right query terms, leading to the retrieval of more irrelevant pages. This has led to the need for the development of new techniques to assist users effectively navigate, trace and organize the available web documents, with the ultimate goal of finding those best matching their needs. Document Clustering is one of the techniques that can play an important role towards the achievement of this objective.

Document clustering has become an increasingly important task in analyzing huge numbers of documents distributed among various sites. Furthermore organizing them into different groups called as clusters, where the documents in each cluster share some common properties according to defined similarity measure. The fast and high-quality document clustering algorithms play an important role in helping users to effectively navigate, summarize, and organize the information.

Document clustering has been studied intensively because of its wide applicability in areas such as Web Mining, Search Engines, Information Retrieval, and Topological Analysis. Document Clustering is different than document classification. In document classification, the classes (and their properties) are known a priori, and documents are assigned to these classes; whereas, in document clustering, the number, properties, or membership (composition) of classes is not known in advance. Thus, classification is an example of supervised machine learning and clustering that of unsupervised machine learning [1]. This distinction is illustrated in figure (1). Document Clustering can produce either disjoint (hard clustering) or overlapping (soft clustering) partitions. In an overlapping partition, it is possible for a document to appear in multiple clusters whereas in disjoint clustering, each document appears in exactly one cluster [2].

Document clustering algorithms may be divided into two groups: Hierarchical algorithms produce a hierarchy of clusters, while Partitioning algorithms give a flat partition of the set.

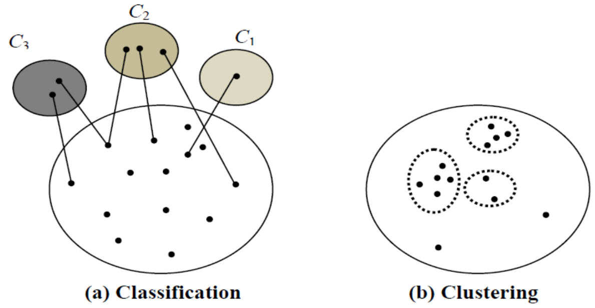


Fig. 1. In (a), three classes are known a priori, and documents are assigned to each of them. In (b), an unknown number of groupings must be inferred from the data based on a similarity criterion [1].

Although standard clustering techniques such as k-means can be applied to document clustering, they usually do not satisfy the special requirements for clustering documents: high dimensionality, high volume of data, ease for browsing, and meaningful cluster labels. In addition, many existing document clustering algorithms require the user to specify the number of clusters as an input parameter [3]-[8]. Incorrect estimation of the value always leads to poor clustering accuracy. Furthermore, many clustering algorithms are not robust enough to handle different types of document sets in a real-world environment. In some document sets, cluster sizes may vary from few to thousands of documents. This variation tremendously reduces the resulting clustering accuracy for some of the state-of-the art algorithms.

The challenges of hierarchical clustering and the weaknesses of the standard clustering methods formulate the need for an accurate, efficient, and scalable clustering method that addresses the special challenges of document clustering. Frequent itemset-based clustering method is shown to be a promising method for high dimensionality clustering in recent literature. It reduces the dimension of a vector space by using only frequent itemsets for clustering. Frequent itemsets form the basis of association rule mining [9]. Exploiting the property of frequent itemsets (each subset of a frequent itemset is also frequent) and using data structures supporting the support counting, the set of all frequent itemsets can be efficiently determined even for large databases. Recent studies on frequent termsets in text mining fall into two categories. One is to use Association Rules to conduct text categorization [10,11] and the other one is to use frequent itemsets for text clustering [12]-[26].

In our prior research [27], we have presented an efficient Association Rules-based Web Document Clustering approach (ARWDC). The main idea of the association rule-based clustering stage is based on a simple observation: the documents under the same topic should share a set of common keywords. Some minimum fraction of documents in the document set must contain these common keywords, and they correspond to the notion of frequent termsets which form the basis of the initial clusters. An essential property of frequent termset is its representation of words that commonly occur together in documents.

To illustrate that this property is important for clustering, we consider two frequent terms, “apple” and “window”. The documents that contain the word “apple” may discuss about fruits or farming. While the documents that contain the word “window” may discuss about renovation. However, if we found association rules between both words occur together in many documents, then we may identify another topic that discusses about operating systems or computers. By precisely identifying these hidden topics as the first step and then clustering documents based on them, we can improve the accuracy of the clustering solution.

The Apriori algorithm remains the most commonly used algorithm in the mining process [9]. The Apriori achieves good reduction on the size of candidate set but still suffers from generating huge numbers of candidates and taking many scans of large databases for frequency checking. Our MTHFT algorithm proposed in [28] for efficient mining of association rules from documents instead of Apriori algorithm. Since by using MTHFT algorithm, the scanning cost and computational cost is improved moreover the performance is considerably increased furthermore increase up the clustering process.

In this paper, we have presented an extensive analysis of the ARWDC approach for different sizes of Reuters datasets. Furthermore the performance of the approach is compared with the existing two clustering algorithms like Bisecting K-means and FIHC and evaluated with the help of evaluation measures such as, Precision, Recall and F-measure.

The organization of the paper is as follows. The concise review of related researches is presented in Section 2. The ARWDC approach based on association rules is described in Section 3. The extensive analysis of the ARWDC approach using different sizes of Reuters datasets moreover the comparison with other clustering algorithms are given in section 4. The conclusion is summed up in Section 5 and the future work in Section 6.

II. Review Of Litruture

In data mining literature, there are limited researches for clustering the data based on association rules mining. Whereas all researches for clustering web documents based on frequent termsets are conducted in web mining field. A review of researches and the work that has been done are presented in this section.

Association Rules Mining is considered the basis of data mining research [9], [29]. The first method of integrating association rules and clustering techniques in an undirected hypergraph is presented in [30]. The frequent itemsets were modeled as hyperedges and a min-cut hypergraph partitioning algorithm was used to cluster items. There has been some theoretical work relating hypergraphs with association rules [31]. Directed hypergraphs [32],[33] extend directed graphs and have been used to model many-to-one, one-to-many and many-to-many relationships in theoretical computer science and operations research.

The method for clustering of data in a high dimensional space based on a hypergraph model is proposed in [34]. In a hypergraph model, each data item represented as a vertex and related data items connected with weighted hyperedges. A hyperedge represented a relationship (affinity) among subsets of data and the weight of the hyperedge reflected the strength of this affinity. A hypergraph partitioning algorithm used to find a partitioning of the vertices such that the corresponding data items in each partition were highly related and the weight of the hyperedges cut by the partitioning minimized. The method is linearly scalable with respect to the number of dimensions of data and items, provided the support threshold used in generating the association rules is sufficiently high. it suffers from the fact that right parameters are necessary to find good clusters.

An algorithm to mine association rules from medical data based on digit sequence and clustering is presented in [35]. The entire database divided into partitions of equal size, each partition called cluster. Each cluster considered one at a time by loading the first cluster into memory and calculating frequent itemsets. Then the second cluster considered similarly and calculating frequent itemsets. This approach reduced main memory requirement since it considered only a small cluster at a time and it is scalable and efficient.

The first criterion for clustering transactions using frequent itemsets, instead of using a distance function is presented in [25]. In principle, this method can also be applied to document clustering by treating a document as a transaction; however, the method does not create a hierarchy for browsing. The novelty of this approach is that it exploits frequent itemsets (by applying Apriori algorithm) for defining a cluster, organizing the cluster hierarchy, and reducing the dimensionality of document sets.

The two clustering algorithms, FTC and HFTC, are proposed in [12]. The basic motivation of FTC is to produce document clusters with overlaps as few as possible. FTC works in a bottom-up fashion. As HFTC greedily picks up the next frequent itemset to minimize the overlapping of the documents that contain both the itemset and some remaining itemsets. The clustering result depends on the order of picking up itemsets, which in turn depends on the greedy heuristic used. The weakness of the HFTC algorithm is that it is not scalable for large document collections.

To measure the cohesiveness of a cluster directly using frequent itemsets, the FIHC algorithm is proposed in [14]. Two kinds of frequent item are defined in FIHC: global frequent item and cluster frequent item. However, FIHC has three disadvantages in practical application: first, it cannot solve cluster conflict when assigning documents to clusters. Second, after a document has been assigned to a cluster, the cluster frequent items were changed and FIHC does not consider this change in afterward overlapping measure. Third, in FIHC, frequent itemsets is used merely in constructing initial clusters.

Frequent Term Set-based Clustering (FTSC) algorithm is introduced in [15]. FTSC algorithm used the frequent feature terms as candidate set and does not cluster document vectors with high dimensions directly. The results of the clustering texts by FTSC algorithm cannot reflect the overlap of text classes. But FTSC and the improvement FTSHC algorithms are comparatively more efficient than K-Means algorithm in the clustering performance.

The document clustering algorithm on the basis of frequent termsets is proposed in [22]. Initially, documents were denoted as per the Vector Space Model and every term is sorted in accordance with their relative frequency. Then frequent term sets can be mined using frequent-pattern growth (FP growth). Lastly, documents were clustered on the basis of these frequent term sets. The approach was efficient for very large databases, and gave a clear explanation of the determined clusters by their frequent term sets. The efficiency and suitability of the proposed algorithm has been demonstrated with the aid of experimental results.

To the best of our knowledge, all previous researchers depend on the frequent termsets for clustering web documents. While we do not know of any research that exploits association rules in web document clustering.

III. Association Rules based Clustering Approach

An effectual approach for clustering a web documents with the aid of association rules is discussed in this section[27]. The ARWDC approach as shown in figure (3) consists of the following major stages:

- Offline Collecting of Documents
- Document Preprocessing
- Association Rules Mining
- Document Clustering
- Post Processing

A. *Offline Collecting of Documents stage*

The first step in the ARWDC approach is collecting and analyzing the documents (i.e. the relevant documents). The process of selecting documents in the ARWDC approach is done offline that means the documents are previously downloaded. The largest Reuters datasets is an example for offline documents [36]. The Reuters-21578 collection is distributed in 22 files. Each of the first 21 files (reut2-000.sgm through reut2-020.sgm) contain 1000 documents, while the last (reut2- 021.sgm) contains 578 documents. Documents were marked up with SGML tags. There are 5 categories Exchanges, Organizations, People, Places and Topics in the Reuters dataset and each category has again sub categories in total 672 sub categories. We have collected the TOPIC category sets to form the dataset. The TOPICS category set contains 135 categories. From these documents we collect the valid text data of each category by extracting the text which is in between <BODY> ,</BODY> and placed in a text document and named it according to the topic.

B. *Document Preprocessing stage*

Preprocessing stage is a very important step since it can affect the result of a clustering algorithm. So it is necessary to pre-process the data sensibly. Preprocessing have the several steps that take a text document as input and output as a set of tokens to be used in feature vector. It begins after collecting the documents that need to be clustered. The ARWDC approach employs several pre-processing steps including stop words removal, stemming on the document set and indexing documents by applying TF*ID:

- Stop words removal: In this process, the documents are filtered by removing the stop-words from documents content and reduce noise. Stop-words are words that from non-linguistic view do not carry information such as (a, an, the, this, that, I, you, she, he, again, almost, before, after). One major property of stop-words is that they are extremely common words.

- Stemming: Removes the prefixes and suffixes in the words and produces the root word known as the stem. Typically, the stemming process will be performed so that the words are transformed into their root form [37]. A good stemmer should be able to convert different syntactic forms of a word into its normalized form, reduce the number of index terms, save memory and storage and may increase the performance of clustering algorithms to some extent; meanwhile it should try stemming. Porter Stemmer [38] is a widely applied method to stem documents. It is compact, simple and relatively accurate. It does not require to create a suffix list before applied. In this paper, we apply Porter Stemmer in our pre-processing .

- Indexing documents: the indexing process has done on the filtered and stemmed documents. The documents indexed automatically by labelling each document by a set of the most important words with their frequencies. The techniques for automated production of indexes associated with documents usually rely on frequency-based weighting schema. The weighting schema is used to index documents and to select the most important words in all document collections. The purpose of weighting schema is to reduce the relative importance of high frequency terms while giving a higher weight value for words that distinguish the documents in a collection. The weighting scheme TF-IDF (Term Frequency, Inverse Document Frequency) is used to assign higher weights to distinguished terms in a document, and it is the most widely used. Weighting scheme is defined as [39]:

$$w(i,j) = tfidf\,(d_i, t_j) = \begin{cases} Nd_i, t_j * \log_2 \dfrac{|C|}{Nt_j} & if\ Nd_i, t_j \geq 1 \\ 0 & if\ Nd_i, t_j = 0 \end{cases} \qquad (1)$$

where w(i,j) ≥ 0, Nd_i, t_j denotes the number the term t_j occurs in the document d_i (term frequency factor), Nt_j denotes the number of documents in collection C in which t_j occurs at least once (document frequency of the term t_j) and │C│ denotes the number of the documents in collection C. The first clause applies for words occurring in the document, whereas for words that do not appear (Nd_i, t_j =0), we set w (i,j)=0. The weighting scheme includes the intuitive presumption that is: the more often a term occurs in a document, the more representative of the content of the document (term frequency). Moreover the more documents the term occurs in, the less discriminating it is (inverse document frequency). Once a weighting scheme has been selected, automated indexing can be performed by simply selecting the words that satisfy the given weight constraints for each document. The major advantage of an automated indexing procedure is that it reduces the cost of the indexing step. For each document, we store all words, with their frequencies and their calculated weighing values. Next, the words that have zero weighted value were eliminated automatically and select only the words that satisfy the given weighting threshold. Finally, the words (the number of words that satisfy the threshold weight value) taken as the final set of words to be used in the Association Rule Mining stage. This is the criteria of using the weight constraints.

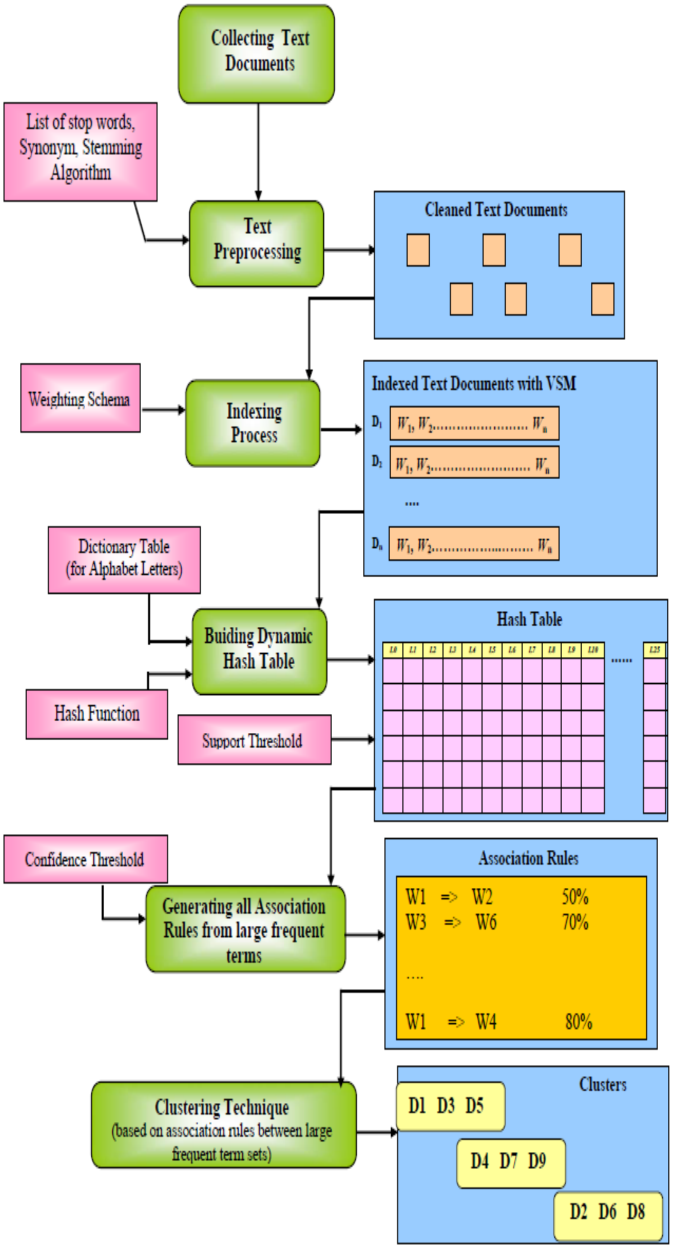

Collecting Text Documents
List of stop words, Synonym, Stemming Algorithm
Text Preprocessing
Cleaned Text Documents
Weighting Schema
Indexing Process
Indexed Text Documents with VSM
D1 W1, W2........................ Wn
D2 W1, W2........................ Wn
....
Dn W1, W2........................ Wn
Dictionary Table (for Alphabet Letters)
Buiding Dynamic Hash Table
Hash Function
Support Threshold
Hash Table
L0 L1 L2 L3 L4 L5 L6 L7 L8 L9 L10 ...... L25
Confidence Threshold
Generating all Association Rules from large frequent terms
Association Rules
W1 => W2 50%
W3 => W6 70%
....
W1 => W4 80%
Clustering Technique (based on association rules between large frequent term sets)
Clusters
D1 D3 D5
D4 D7 D9
D2 D6 D8

Fig. 2. ARWDC approach.

C. Association Rules Mining Stage

Association rules can be used to solve the problem of finding clusters of similar items. For instance, in market-basket type data, a practical application of association rules is to identify clusters of similar items based on the customer sales information. This helps to understand patterns in sales of items and to group items based on customer interests. Association rule mining is to find out association rules that satisfy the predefined minimum support and confidence from a given database. The problem is usually decomposed into two sub-problems: 1) One is to find those itemsets whose occurrences exceed a predefined threshold in the database; those itemsets are called frequent or large itemsets, 2) The second problem is to generate association rules from those large itemsets with the constraints of minimal confidence.

Apriori algorithm considered to be the basic for all developed algorithms to solve the first problem. However there are two drawbacks of the Apriori algorithm. One is the complex candidate generation process that uses most of the time, space and memory. Another drawback is the multiple scan of the database. Although the drawbacks of the Apriori algorithm, it still use for generating the frequent termsets that used in the document clustering. In order to speed up the mining process as well as to address the scalability with different documents regardless of their sizes, we used our algorithm [28] called Multi-Tire Hashing Frequent Termsets algorithm (MTHFT) in figure (4) to generate all strong association rules. It is basically different from all the previous algorithms since it overcomes the drawbacks of Apriori algorithm by employing the power of data structure called Multi-Tire Hash Table. Moreover it uses new methodology for generating frequent termsets by building the hash table during the scanning of documents only one time consequently, the number of scanning on documents decreased.

Once the frequent termsets from documents have been generated, it is straightforward to generate all strong association rules from them (where strong association rules satisfy both minimum support and minimum confidence). This can be done using the following equation for confidence, where the conditional probability is expressed in terms of termsets support count [40] :

$$Confidence\ (A \rightarrow B) = p(B \backslash A) = \frac{support_{count}(A \cup B)}{support_{count}(A)} \qquad (2)$$

where $port_{count}$ $(A \cup B)$ is the number of documents containing the termsets $(A \cup B)$, and $support_{count}$ *(A)* is the number of documents containing the termset *A*.

1) The advantages of MTHFT Algorithm: The MTHFT algorithm has many advantages summarized as follows:

- Provides facilities to avoid unnecessary scans to the documents, which minimize the I/O. Where the scanning process occurs on the hash table instead of whole documents compared to Apriori algorithm
- The easy manipulations on hash data structure and directly computing frequent termsets are the added advantages of this algorithm, moreover the fast access and search of data with efficiency.
- MTHFT shows better performance in terms of time taken to generate frequent termsets when compared to Apriori algorithm. Furthermore, it permits the end user to change the threshold support and confidence factor without re-scanning the original documents since the algorithm saves the hash table into secondary storage media.
- The main advantage of this algorithm is that, it is scalable with all types of documents regardless of their sizes.
- Depending on the multi-tire technique in building the primary bucket, each bucket can store only a single element then we cannot associate more than one term with a single bucket, which is a problem in the case of collisions.

MTHFT Algorithm:
T_m: Set of all termsets for each document *d*
C_m: Candidate termsets for each document *d*
I_k : Frequent termsets of size *k*.
AR_k: Association Rules of size *k*

Input: All Text documents.
Process logic: Building Multi-Tire Hash Table and Finding the frequent termsets.
Output: Generating all strong Association Rules.

```
for each document $d_m \in D$ do begin
    $T_m = \{ t_i : t_i \in d_m ,\ 1 \le i \le n \}$
        for each term $t_i \in T_m$ do
            $h(t_i) = t_i \bmod N;$
            $t_i.count{+}{+};$
                // insert each term in hash table
        end
        $C_k$ = all combinations of $t_i \in d_m$
        $C_m = subset(C_k , d_m);$
            for each candidate $c_j \in C_m$ do
                $h(c_j) = c_j \bmod N;$
                $c_j.count{+}{+};$
            // insert each candidate in hash table
            end
end
  for given s= minsup in hash table do
      $I_1 = \{t \mid t.count \ge minsup \}$
      $I_k = \{c \mid c.count \ge minsup, k \ge 2\}$
  end
      for given c= minconf in $I_k$ do
          $AR_k = \{ I_i \rightarrow I_j \mid confidence \ge minconf, k \ge 2\}$
  end
```

Fig. 3. The MTHFT algorithm.

D. Documents Clustering Stage

Document clustering algorithm based on association rules considered a keyword-based algorithm which picks up the core rules between words with specific criteria and groups the documents based on these keywords. This approach includes five main steps:

- Picking out all strong Association Rules
- Constructing initial partitions

- Merging Similar Partitions
- Making Partition Disjoint
- Clustering Documents

1) Picking out all Strong Association Rules: The Multi-Tire Hashing Frequent Termsets algorithm is used in the previous step to find out all strong association rules furthermore to speeding up the mining process. It have ability to determine large frequent termsets at different minimum support threshold values without redoing the mining process again. Therefore, we can generating different sets of association rules between different frequent termsets in the clustering process easily. We start with a set of association rules R_s *generated between the set of 2-large frequent termsets s since* $R_k = I_i \rightarrow I_j$.

$$R_s = \{ R_1, R_2, R_3, \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots, R_k \} \qquad (3)$$

2) Constructing Initial Partitions: initially, we sort the set of all strong association rules R_s *in descending order in accordance with their confidence level as in (4):*

$$Conf(R_1) > Conf(R_2) > \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots Conf(R_k) \qquad (4)$$

An initial partition P_1 is constructed for first association rule in R_s. Afterward, all the documents containing both termsets that constructed the rules are included in the same cluster. Next, we take the second association rules whose confidence is less than the previous one to form a new partition P_2. This partition is formed by the same way of the partition P_1. This procedure is repeated until every association rules moved into partition P_i since

$$P_i = < R_i\ ,\ doc\ [\ R_i] > \qquad (5)$$

Since a document usually contains more than one frequent termset, the same document may appear in multiple initial partitions, i.e., initial partitions are overlapping. The purpose of initial partitions is to ensure the property that all the documents in a cluster contain all the terms in the association rules that defines the partition. These rules can be considered as the mandatory identifiers for every document in the partition. We use these association rules as the partition label to identify the partition . The main purpose of presenting the partition label is to facilitate browsing for the user.

3) Merging Similar Partitions: in this step, all partitions that contain the similar documents are merged into one partition. The benefit of this step is reducing the number of resulted partitions.

4) Making partitions Disjoint: in this step, we remove the overlapping of partitions since there are some documents belong to one or more initial partitions. we assign a document to the "Optimal" initial partition so that each document belongs to exactly one partition. This step also guarantees that every document in the partition still contains the mandatory identifiers. We propose the Weighted Score ($P_i \leftarrow doc_j$*) in equation (6) to measure the optimal initial partition* P_i *for a document* doc_j.

$$\left(\boldsymbol{P_i} \leftarrow \boldsymbol{doc}_{\boldsymbol{j}}\right) = \sum_{\boldsymbol{k}} \boldsymbol{w_k} * \boldsymbol{m_i}/\boldsymbol{n_w} \qquad (\boldsymbol{6})$$

where $\sum_k w_k$ represents the sum of weighted values of all words constructed the association rules from doc_j , m_i represents the number of documents in the initial partition P_i, and n_w represents the number of words that construct the partition P_i from doc_j. The weighted values of words w_k are defined by the standard inverse document frequency (TF-IDF) in the indexing process in section (III.B). The Weighted Score measure used the weighed values of frequent termsets instead of the number of occurrences of the terms in a document. Since the weighted values are an important piece of information based on the intuitive presumption of the weighting schema that is: the more often a term occurs in a document, the more representative of the content of the document (term frequency). Moreover the more documents the term occurs in, the less discriminating it is (inverse document frequency). To make partitions non-overlapping, we assign each doc_j to the initial partition P_i of the highest $score_i$. After this assignment, if there are more than one P_i that maximizes the Weighted Score $\left(P_i \leftarrow doc_j\right)$, we will choose the one that has the most number of words in the partition label. After this step, each document belongs to exactly one partition.

Example**:** Consider we have eleven documents to do clustering process. They are manually selected from different four topics (*Economy, Computer Science, Sports*, and *Avain Bird Flue*). Each document is indexed by a set of weighted words. After the mining process, we generated a set of strong association rules from 2-large frequent termsets equalls to 226 rule with 50% minimum confidence. The initial partitions of this example are constructed equals to 131 partition. After merging partitions based on the the similar documents we have 15 partition as shown in Table 1.

From the table, we observed that there are more than one document belongs to more than one partition *for example*, D_7 belongs to (P_1, P_3 and P_{15}) and D_5 belongs to (P_{10} , and P_{11}) and so on. To remove the overlapping between partitions and find the optimal partition for a document doc_j, we need to calculate its scores against each initial partition that contains the document as follows: to find the optimal partition for D_7 so that we begin to calcuate its scors against each initial partition (P_1 , P_3 and P_{15})

Weighted Score ($P_1 \leftarrow D_7$)

= (2.45+1.87+2.45+4.91+2.45+4.91) * 2 / 6

= 6.34

Weighted Score ($P_3 \leftarrow D_7$)

= (2.45+2.45+2.45+4.91+2.45+4.91) * 1 / 6

= 3.27

Weighted Score ($P_{15} \leftarrow D_7$)

= (2.45+1.87) * 2 / 2 = 4.32

TABLE I. INITIAL PARTITIONS

Initial Partitions	Text Documents
P_1	D$_8$,**D$_7$**
P_2	D$_9$,D$_{10}$,D$_{11}$
P_3	**D$_7$**
P_4	D$_9$,D$_{11}$
P_5	D$_8$
P_6	D$_1$
P_7	D$_1$,D$_2$
P_8	D$_6$
P_9	D$_4$
P_{10}	D$_5$
P_{11}	D$_4$,D$_5$
P_{12}	D$_6$,D$_8$
P_{13}	D$_1$,D$_2$,D$_3$
P_{14}	D$_1$,D$_3$
P_{15}	D$_6$,**D$_7$**

From the above calculation, D$_7$ will assign to P_1 which has the highest score. After repeating the above computation for each document, each document belongs to exactly one partition as shown in Table 2.

5) *Clustering Documents:* after removing the overlapping and put each document in its optimal partition, we begin to clustering documents based on the partition labels. In this step, we don't require to pre-specified number of clusters as previous standard clustering algorithms. we have a set of non-overlapping partitions P_i and each partition has a number of documents D_j. We first identify the association rules that construcr each partition. The set of all words that construct all association rule in P_i called the labeling Words Ld $[W_i]$. Moreover every document in the partition must contain all the words in the partition label. We use the partition label to identify the partition.

TABLE II. DISJOINT PARTITIONS

Initial Partitions	Text Documents
P_1	D$_8$,D$_7$
P_2	D$_9$,D$_{10}$,D$_{11}$
P_3	D$_4$,D$_5$
P_4	D$_6$
P_5	D$_1$,D$_2$,D$_3$

We observed that the partition labeling words based on association rules are more informative than other based on frequent termsets in [28]. However the number of association rules always greater than the number of frequent termsets, the rules carry out more information and identify hidden knowledge from documents help us to improve the accuracy of the clustering process.

The definition of the similarity measure plays an important role in obtaining effective and meaningful clusters. For each document $D_j$in partition P_i, to compute its similarity measure we must obtain the Derived keywords Vd $[W_i]$ from taking into account the difference words between the top weighted frequent words for each document with the labeling words. Subsequently the total support of each derived word is computed within the partition. The set of words satisfying the partition threshold (the percentage of the documents in partition P_i that contains the termset) are formed as Descriptive Words Pw $[C_i]$ of the partition P_i. Afterward, we compute the similarity of each document in the partitions with respect to the descriptive words. The similarity between two documents S$_m$ is computed as in [41]. Based on the similarity measure, a new cluster is formed from the partitions i.e. each cluster will contain all partitions that have the similar similarity measures.

E. Post processing

For different applications there are different ways to do post processing. One common post processing is to select a suitable threshold to generate the final cluster result. After document clustering we get a basic cluster map in which the clusters are organized like a tree or in a flat way. Thereby some post processing algorithms may be applied to find out the correct clusters relation.

IV. EXPERIMENTAL RESULTS AND PERFORMANCE EVALUATION

Our experiments have been performed on a personal computer with a 2.50 GHz CPU and 6.00 GB RAM and we chose the programming language C#.net for the implementation because it allows fast and flexible development. The largest dataset, Reuters, is chosen to exam the efficiency and scalability of the ARWDC approach. To evaluate the effectiveness of the ARWDC approach, this section presents the result comparisons with some of the popular hierarchical document clustering algorithms like Bisecting K-means and FIHC for clustering web documents. The rest of this section first explains the evaluation measures, and finally presents and analyzes the experiment results.

A. Evaluation Methods

The F-measure, as the commonly used external measurement, is used to evaluate the accuracy of our clustering algorithms. F-measure is an aggregation of Precision and Recall concept of information retrieval. Recall is the ratio of the number of relevant documents retrieved for a query to the total number of relevant documents in the entire collection as in (7):

$$Recall\,(K_i\,,C_j) = \frac{n_{ij}}{|K_i|} \tag{7}$$

Precision is the ratio of the number of relevant documents to the total number of documents retrieved for a query as in (8):

$$Precision\,(K_i\,,C_j) = \frac{n_{ij}}{|C_j|} \tag{8}$$

while F-measure for cluster $C_{\mathrm{j}}$and class K_{i} is calculated as in (9):

$$F\,(K_i\,,C_j) = \frac{2 * Recall\,(K_i\,,C_j) * Precision\,(K_i\,,C_j)}{Recall\,(K_i\,,C_j) + Precision\,(K_i\,,C_j)} \tag{9}$$

where n_{ij} is the number of members of class K_i in cluster C_j. $|C_j|$ is the number of members of cluster C_j and $|K_i|$ is the number of members of class K_i .

The weighted sum of all maximum F-measures for all natural classes is used to measure the quality of a clustering result *C*. This measure is called the *overall F-measure* of *C*, denoted $F(C)$ is calculated as in (10):

$$F(C) = \sum_{K_i \in K} \frac{|K_i|}{|D|} max_{C_j \in C}\{F(K_i, C_j)\} \qquad (10)$$

where *K* denotes all natural classes; *C* denotes all clusters at all levels; $|K_i|$ denotes the number of documents in natural class K_i; and $|D|$ denotes the total number of documents in the dataset. The range of $F(C)$ is [0,1]. A large $F(C)$ value indicates a higher accuracy of clustering.

B. *Experimental Results*

In this section, we evaluate the performance of the ARWDC approach in terms of the efficiency, accuracy and scalability compared to Bisecting K-means and FIHC algorithms. We chose Bisecting k-means because it has been reported to produce a better clustering result consistently compared to k-means and agglomerative hierarchical clustering algorithms. FIHC is also chosen because it uses frequent word sets. For a fair comparison, we did not implement Bisecting k-means and FIHC algorithms by ourselves. We downloaded the CLUTO toolkit [42] to perform Bisecting k-means, and obtained FIHC [43] from their author.

- ***Performance Investigations on Accuracy***

The F-measure represents the clustering accuracy. Table 3 shows the F-measure values for all three algorithms with different user specified numbers of clusters. Since ARWDC and HFTC do not take the number of clusters as an input parameter, we use the same minimum support 15% in Reuters dataset to ensure fair comparison.

From table (3), The highlighted results show that our ARWDC approach is better than Bisecting k-means and FIHC algorithms for specified Reuters data set. Furthermore the final average results indicate that the ARWDC outperforms all other algorithms in accuracy for most number of clusters.

Fig. 4 shows the comparison between all the three clustering approaches based on the overall F-measure values with different numbers of clusters. It illustrates that the ARWDC has the higher F-measure values than all competitive algorithms because it uses a better model for text documents. Higher F-measure shows the higher accuracy.

- ***Performance Investigations on Efficiency and Scalability***

The largest dataset, Reuters, is chosen to exam the efficiency and scalability of our approach. Many experiments were conducted to exam the efficiency of ARWDC approach.

TABLE III. F-MEASURE COMPARISON OF CLUSTERING ALGORITHMS

Datasets	# of	Overall F-measure		
	Clusters	Bisecting k-means	FIHC	ARWDC
Reuters 21578	3	0.34	0.53	**0.57**
	15	0.38	0.45	**0.56**
	30	0.38	0.43	**0.53**
	60	0.27	0.38	**0.59**
	average	0.41	0.44	**0.55**

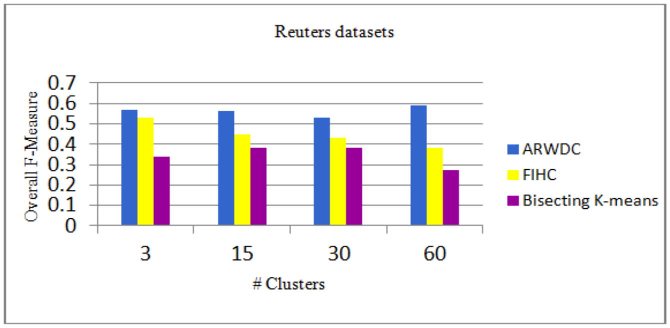


Fig. 4. Overall F-measure results comparison with Reuters dataset.

Figure 5 compares the runtime of ARWDC with bisecting k-means and FIHC algorithms on different sizes of documents of Reuters. The minimum support is set to 15% to ensure that the accuracy of all produced clustering are approximately the same. The number of documents is taken as X-axis and the time taken to find the clusters is taken as Y-axis. ARWDC approach runs approximately twice faster than the others. This is returned to the effect of using MTHFT algorithm for mining association rules. Since the execution time is decreased to mine association rules as support decreased in compared to Apriori algorithm. We conclude that ARWDC is more efficient than other approaches.

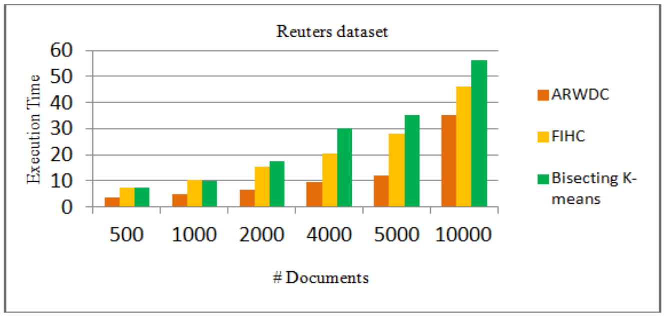


Fig. 5. Efficiency comparison of ARWDC with FIHC and Bisecting *K*-means on different sizes of Reuters at minsup=15%.

A large dataset from Reuters are created for examining the scalability of ARWDC approach. We duplicated the files in Reuters until we get 20000 documents. Figure 6 illustrates that our algorithm runs approximately twice faster than bisecting k-means and FIHC in this scaled up document set.

Figure 7 and 8 illustrate the runtimes with respect to the number of documents for different stages of AREDC approach and FIHC algorithm. Figure 7 shows that the MTHFT and clustering are not time-consuming stages since MTHFT algorithm improved the mining process and speed up the clustering stage. It demonstrates that ARWDC is a very scalable method.

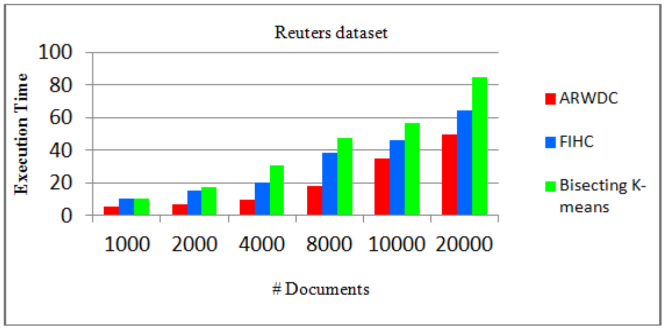


Fig. 6. Scalability comparison of ARWDC, FIHC and Bisecting K-means with scale up document set.

Figure 8 also shows that the Apriori and the clustering are the most time-consuming stages in FIHC, while the runtimes of MTHFT and clustering stages are comparatively short. Since the efficiency of the Apriori is very sensitive to the input parameter minimum support. Consequently, the runtime of FIHC is inversely related to this parameter. In other words, runtime increases as minimum support decreases.

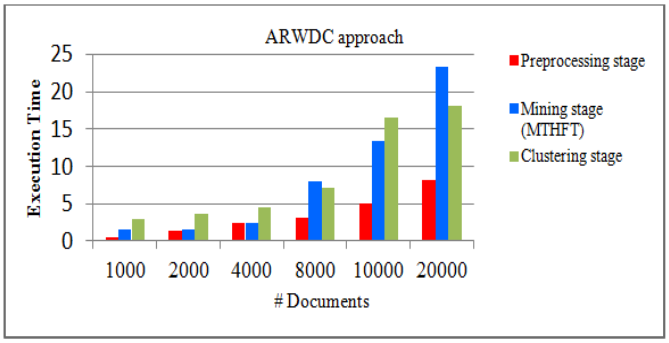


Fig. 7. Scalability comparison of ARWDC approach on different sizes of Reuters for all different stages.

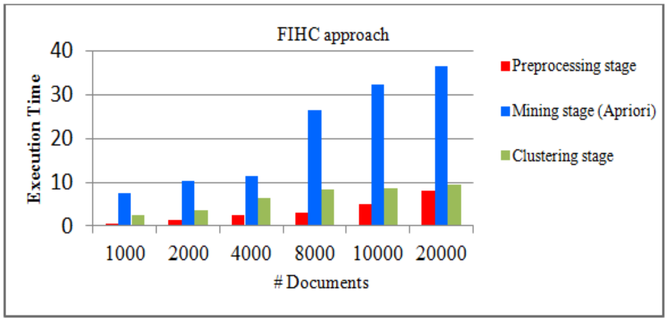


Fig. 8. Scalability comparison of FIHC algorithm on different sizes of Reuters for all different stages.

In conclusion, the major advantages of our ARWDC approach are as follows:

- By generating the strong association rules with specific criteria , the dimensionality of a document is drastically reduced. This is a key factor for the efficiency and scalability of ARWDC approach.
- Experimental results show that ARWDC outperforms the well-known clustering algorithms in terms of accuracy. It is robust and consistent even when it is applied to large and complicated document sets.
- Many existing clustering algorithms require the user to specify the desired number of clusters as an input parameter. ARWDC treats it only as an optional input parameter. Close to optimal clustering quality can be achieved even when this value is unknown.
- Easy to browse with more informative and meaningful partition labels since each partition has a set of association rules which a user may utilize for browsing.
- Since a real world document set may contain a few hundred thousand of documents, experiments show that our approach is significantly more efficient and scalable than all of the tested competitors.

II. Conclusion

In this paper, we have conducted an extensive analysis of association rules-based web document clustering ARWDC approach. The largest dataset, Reuters, is chosen to exam the efficiency and scalability of our algorithm. The experimental results show that at different sizes of Reuters datasets, the ARWDC approach improved scalability. Furthermore when compared with other clustering algorithms like Bisecting K-means and FIHC, the accuracy and efficiency are improved. Moreover, ARWDC approach associated a meaningful label to each final cluster. Then the user can easily find out what the cluster is about since the label can provide an adequate description of the cluster based on Association Rules. However, it is time-consuming to determine the labels after the clustering process is finished. From all experiments, we conclude that ARWDC approach has favorable quality in clustering documents using Association Rules.

III. Future Work

The importance of document clustering will continue to grow along with the massive volumes of web documents. With the standardization of XML as an information exchange language over the web, documents formatted in XML have become quite popular. Moreover, most of the clustering algorithms of MEDLINE abstracts are based on pre-defined categories. In future, we intend to apply ARWDC approach for automatically clustering the MEDLINE abstracts formatted in XML to help biomedical researchers in quickly finding relevant and important articles related to their research field without need to predefine categories.

References

[1] N. Andrews, and E. Fox, “Recent developments in document clustering,” Technical Report TR-07-35, Computer Science, Virginia Tech. April 2007.

[2] S. Sharma, and V. Gupta, “Recent developments in text clustering techniques,” in Proc. of Int. J. of Computer Applications, vol. 37, pp. 14-19, 2012.

[3] K. Jain, N. Murty, and J. Flynn, “Data clustering: a review,” in Proc. of Int. Conf. on ACM Computing Surveys, vol. 31, pp. 264-323, 1999.

[4] M. Steinbach, G. Karypis, and V. Kumar, “A comparison of document clustering techniques,” KDD Workshop on Text Mining, 2000. Avaiable online : http://glaros.dtc.umn.edu/gkhome/node/157

[5] P. Berkhin, (2004)“Survey of clustering data mining techniques,” [Online]. Available: http://www.accrue.com/products/rp_cluster_review pdf

[6] R. Xu, “Survey of clustering algorithms,” in Proc. of Int. Conf. of IEEE Transactions on Neural Networks, vol. 15, pp. 634-678, 2005.

[7] F. Benjamin, W. Ke, and E. Martin, “Hierarchical document clustering,” Simon Fraser University, Canada, pp. 555-559, 2005.

[8] J. Ashish, and J. Nitin, “Hierarchical document clustering: A review,” in Proc. of 2nd National. Conf. on Information and Communication Technology, 2011, Proceedings published in Int. J. of Computer Applications.

[9] R. Agrawal, T. Imielinski, and A. Swami, “Mining association rules between sets of items in large databases,” in Proc. of Int. Conf. on Management of Data, vol. 22, pp. 207-216, Washington 1993.

[10] O. Zaiane and M. Antonie, “Classifying text documents by association terms with text categories,” in Proc. of Int. Conf. of Australasian Database, vol. 24, pp. 215-222, 2002.

[11] B. Liu, W. Hsu and Y. Ma, “Integrating classification and association rule mining,” in Proc. of Int. Conf. of ACM SIGKDD on Knowledge Discovery and Data Mining, pp. 80-86, 1998.

[12] M. Beil, and X. Xu, “Frequent term-based text clustering,” in Proc. of Int. Conf. on Knowledge Discovery and Data Mining, pp. 436- 442, 2002.

[13] O. Zamir and O. Etzioni, “Web document clustering: A feasability demonstration,” in Proc. of Int. Conf. of ACM SIGIR, pp. 46-54, 1998.

[14] M. Hassan and K. John, “High quality, efficient hierarchical document clustering using closed interesting itemsets,” in Proc. of Int. IEEE Conf. of on Data Mining, pp. 991-996, 2006.

[15] L. Xiangwei, H. Pilian, A study on text clustering algorithms based on frequent term sets, Springer-Verlag Berlin Heidelberg, 2005.

[16] Y.J. Li, S.M. Chung, J.D. Holt, “Text document clustering based on frequent word meaning sequences,” in Proc. of Int. Conf. of Data & Knowledge Engineering, vol. 64, pp. 381–404, 2008.

[17] H. Edith, A.G. Rene, J.A. Carrasco-Ochoa, and J.F. Martinez-Trinidad, “Document clustering based on maximal frequent sequence,” in Proc. of Int. Conf. of FinTal, vol. 4139, pp. 257-267, LNAI 2006.

[18] Z. Chong, L. Yansheng, Z. Lei and H. Rong, FICW: Frequent itemset based text clustering with window constraint, Wuhan University journal of natural sciences, vol. 11, pp. 1345-1351, 2006.

[19] L. Wang, L. Tian, Y. Jia and W. Han, “A Hybrid algorithm for web document clustering based on frequent term sets and k-means,” Lecture Notes in Computer Science, Springer Berlin, vol. 4537, pp. 198-203, 2010.

[20] Z. Su, W. Song, M. Lin, and J. Li, “Web text clustering for personalized e-Learning based on maximal frequent itemsets,” in Proc. of Int. Conf. on Computer Science and Software Engineering, vol. 06, pp. 452-455, 2008.

[21] Y. Wang, Y. Jia and S. Yang, “Short documents clustering in very large text databases,” Lecture Notes in Computer Science, Springer Berlin, vol. 4256, pp. 38-93, 2006.

[22] W. Liu and X. Zheng, “Documents clustering based on frequent term sets,” in Proc. of Int. Conf. on Intelligent Systems and Control, 2005.

[23] H. Anaya, A. Pons and R. Berlanga, “A Document clustering algorithm for discovering and describing topics,” Pattern Recognition Letters, vol. 31, pp. 502-510, April 2010.

[24] R. Kiran, S. Ravi, and p. Vikram, “Frequent itemset based hierarchical document clustering ung Wikipedia as external knowledge,” Springer-Verlag Berlin Heidelberg, 2010.

[25] B. Fung, K. Wang, and M. Ester, “Hierarchical document clustering using frequent itemsets,” in Proc. of Int. Conf. on Data Mining, vol. 30, pp. 59-70, 2003.

[26] R. Baghel and Dr. R. Dhir, A Frequent concept based document clustering algorithm, in Proc. of Int. J. of Computer Applications, vol. 4, pp. 0975 – 8887, 2010.

[27] N. Negm, P. Elkafrawy, and A. Salem, “An Efficient hash-based association rule mining approach for document clustering,” in Proc. of Int. 16th WSEAS Conf. on COMPUTERS, July 14-17, pp. 376-381, 2012, Kos Island, Greece.

[28] N. Negm, P. Elkafrawy, M. Amin, and A. Salem, “Clustering web documents based on efficient multi-tire hashing algorithm for mining frequent termsets,” in Proc. of Int. J. of Advanced Computer Science and Applications, vol. 2, pp. 6-14, 2013.

[29] R. Agrawal and R. Srikant, “Fast algorithms for mining association rules,” in Proc. of Int. Conf. of Very Large Data Bases, VLDB, pp. 487–499. Morgan Kaufmann, 12–15, 1994.

[30] E. Han, G. Karypis, V. Kumar, and B. Mobasher, “Clustering based on association rule hypergraphs,” in Proc. of SIGMOD Workshop Research Issues on Data Mining and Knowledge Discovery(DMKD ’97), 1997.

[31] D. Gunopulos, H. Mannila, R. Khardon, and H. Toivonen, “Data mining, hypergraph transversals, and machine learning (extended abstract),” in Proc. of Int. Conf. on PODS, pp. 209–216, 1997.

[32] G. Italiano, G. Ausiello, and U. Nanni, “Dynamic maintenance of directed hypergraphs,” in Proc. of Int. Conf. on Theoretical Computer Science, 72(2-3), pp.97–117, 1990.

[33] G. Gallo, G. Longo, and S. Pallottino, “Directed hypergraphs and applications,” in Proc. of Int. Conf. on Discrete Applied Mathematics, 42(2), pp.177–201, 1993.

[34] E.Han, G. Karypis, V. Kumar, and B. Mobasher, “ Hypergraph based clustering in high dimensional data sets :A Summary of Results,” Copyright 1997, IEEE.

[35] M. Jabbar, P. Chandra, and B. Deekshatulu, Cluster based association rule mining for heart attack prediction, in Proc. of J. of Theoretical and Applied Information Technology, 31st October 2011, vol. 32, no. 2, pp.196-201, 2011.

[36] http://kdd.ics.uci.edu/databases/reuters21578/reuters21578.html

[37] J. Jayabharathy, S. Kanmani, and A. Ayeshaa Parveen, A survey od document clustering algorithm with topic discovery, in Proc. of J. of Computing, February 2011, vol. 3, no. 2.

[38] http://tartarus.org/martin/PorterStemmer/

[39] M. Berry, Survey of text mining: clustering, classification, and retrieval,” Springer-Verlag New York, Inc., 2004.

[40] B. Liu, Web Data Mining. Exploring Hyperlinks, Contents, and Usage Data, 2nd ed, Springer-Verlag Berlin Heidelberg, 2011.

[41] S. Krishna, and S. Bhavani, An ffficient approach for text clustering based on frequent itemsets, in Proc. of European J. of Scientific Research, vol.42, no.3, pp.399-410, 2010.

[42] http://glaros.dtc.umn.edu/gkhome/views/cluto

[43] http://ddm.cs.sfu.ca/dmsoft/Clustering/fihc_index.html

Directed and Almost-Directed Flow Loops in Real Networks

M.Todinov
Department of Mechanical Engineering and Mathematical Sciences
Oxford Brookes University
Oxford, Wheatley, OX33 1HX, UK

***Abstract*—Directed flow loops are highly undesirable because they are associated with wastage of energy for maintaining them and entail big losses to the world economy. It is shown that directed flow loops may appear in networks even if the dispatched commodity does not physically travel along a closed contour. Consequently, a theorem giving the necessary and sufficient condition of a directed flow loop on randomly oriented straight-line flow paths has been formulated and a close-form expression has been derived for the probability of a directed flow loop. The results show that even for a relatively small number of intersecting flow paths, the probability of a directed flow loop is very large, which means that the existence of directed flow loops in real networks is practically inevitable. Consequently, a theorem and an efficient algorithm have been proposed related to discovering and removing directed flow loops in a network with feasible flows.**

The new concept 'almost-directed flow loop' has also been introduced for the first time. It is shown that the removal of an almost-directed flow loop also results in a significant decrease of the losses. It is also shown that if no directed flow loops exist in the network, the removal of an almost-directed flow loop cannot create a directed flow loop.

***Keywords*—directed flow loops; almost-directed flow loops; flow networks; optimization; classical algorithms; maximising the flow.**

I. Directed Loops Of Flow In Networks

The existence of routing loops have already been reported in computer networks [1,2]. Due to inconsistencies in routing state among a set of routers, the packets physically travel along a closed loop and never reach their destination. Surprisingly, directed loops of commodity may exist even if none of the dispatched commodities physically travels along a closed loop. This point is illustrated by the examples in Fig.1 featuring supply networks (e.g. supply of petrol from a number of fuel terminals to a number of filling stations), where the same exchangeable commodity is transported along straight lines which are the shortest paths from sources to destinations. Selecting the shortest paths for a data transfer for example, is also a common strategy in communication networks [3].

Suppose that the throughput capacity of each source-destination straight-line path is 10 units. Despite that none of the dispatched commodities physically travels along a closed contour, a directed loop carrying 10 units of flow effectively appears between the intersection points (real or imaginary) x1,x2 and x3 in the network from Fig.1a and between nodes x1,x2,x3 and x4 in the network from Fig.1b.

Removing 10 units of flow from the segments (x1,x2), (x2,x3) and (x3,x4) in Fig.1a and from the segments x1,x2), (x2,x3), (x3,x4) and (x4,x1) in Fig.1b turns the flow circulating along the contours x1,x2,x3,x1 and x1,x2,x3,x4,x1 into zero, without diminishing the amount of total flow sent from the source nodes to destination nodes (Fig.1c,1d).

Figure 1 shows that directed loops of flow can even be found in networks where the intersecting source-destination paths are straight-line segments and no transported commodity physically travels along a closed contour.

A closed contour formed by a sequence of n nonempty sections ($n \geq 3$), in which the flows point along the direction of traversing the contour will be referred to as "directed flow loop".

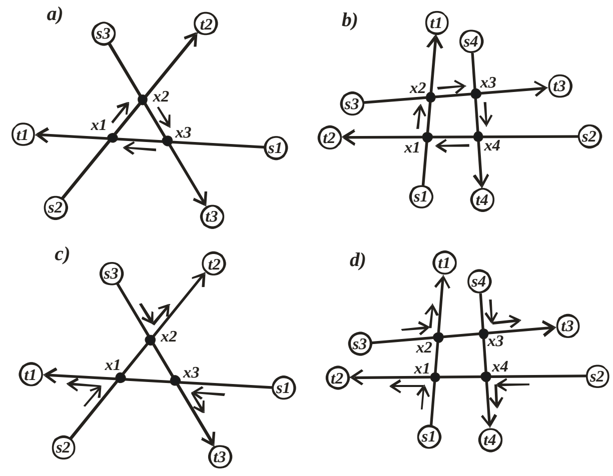


Fig. 1. a,b) Directed closed flow loops naturally appear in networks where the same type of commodity is transported between source-destination pairs; c,d) The directed flow loops can be removed without affecting the throughput flow from sources to destinations.

The directed loops of flow are highly undesirable because: (i) they increase unnecessarily the cost of transportation of the flow in the network, (ii) they consume residual capacity from the edges of the network and (iii) energy is unnecessarily wasted for maintaining the directed flow loops. The presence of directed loops of flow in networks causes big financial losses in the affected sectors of the economy. In computer networks, directed loops of flow consume bandwidth capacity unnecessarily, increase data traffic and ultimately lead to congestion and delayed data transmission. This affects negatively the quality of service of the network.

In supply networks, the existence of directed loops of flow means high transportation costs because energy is wasted on circulating commodities unnecessarily.

The probability of existence of a directed flow loop between the intersection points of random source-destination paths has not yet been considered in the literature, despite its importance. Finding the strongly connected components of a graph, which implies the existence of cyclic paths has been has been considered before [4]. The question of identifying and removing directed flow loops in flow networks however, has been evading the attention of researchers for a very long time. This is evidenced by the fact that in spite of the years of intensive research on static flow networks, the algorithms for maximising the throughput flow published since 1956 leave highly undesirable directed loops of flow in the “optimised” networks. This surprising omission has already been demonstrated in [5] and [6].

There have been a number of published algorithms for optimising the flows in networks. Research related to optimizing network flows has been reviewed in [7-16]. Most of this research is related to determining the edge flows which maximise the throughput flow transmitted from a number of sources to a number of destinations (sinks).

There are two main categories of algorithms solving this problem. The augmentation algorithms preserve the feasibility of the network flow at all steps, until the maximum throughput flow is attained [17-20].

The second major category of algorithms for optimising the throughput flow are based on the preflow concept proposed in [21] and subsequently used as a basis for the algorithms proposed in [22] and [23]. For the preflow, the sum of all edge flows going into a node is allowed to exceed the sum of all flows going out of the node. As a consequence, the flow conservation law at the nodes may be violated and the nodes may contain excess flow. The central idea behind the preflow-push algorithms is converting the preflow into a feasible flow.

In a recent work [6], it was shown that optimising the network flow by using classical augmentation and preflow-push algorithms does not guarantee that there will be no directed flow loops in the optimised networks.

This point can be illustrated immediately with Fig.2, featuring a flow network with three sources s1, s2 and s3, each with capacity 100 units of flow per unit time and three destinations (sinks) t1,t2 and t3, each with capacity 100 units of flow per unit time. Suppose, for the sake of simplicity that the capacities of the separate connecting edges are also 100 units of flow per unit time. To maximise the throughput flow from the sources to the sinks, the classical Edmonds and Karp shortest-path algorithm [18] proceeds with saturating the shortest path (1,2,3,4) with 100 units of flow, followed by saturating the next shortest path (5,6,7,3,8,9,10) with 100 units of flow and finally, with saturating the remaining path (11,12,13,14,15,8,2,16,17,18) with 100 units of flow. As a result, a directed flow loop (2, 3, 8, 2) appears, carrying 100 units of flow. This flow loop is not only associated with wastage of energy. It also congests the network and makes it impossible to transfer additional flow, for example, from node 8 to node 2.

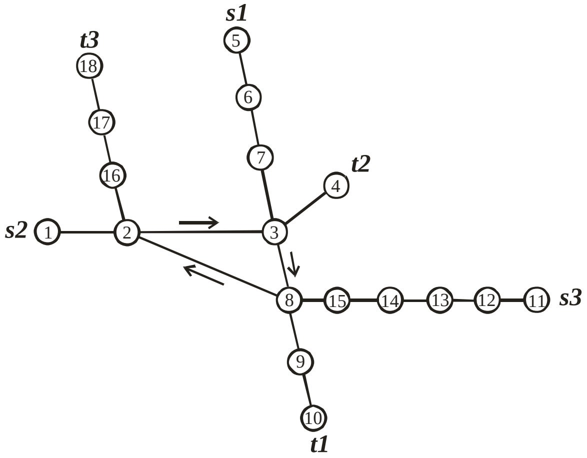


Fig. 2. Network, demonstrating that selecting sequentially the shortest paths between sources and destinations leaves a directed flow loop (2,3,8,2) in the optimised network. All edges have a flow capacity of 100 units.

Finally, to the best of our knowledge, no published analyses exist on almost-directed flow loops which are also associated with losses. The almost-directed flow loops are introduced and defined rigorously in the next section.

Consequently, the objectives of this paper are:

a) To show that directed flow loops can exist in networks even if all none of the dispatched commodity physically travels along a closed contour.

b) To estimate precisely the probability of a directed flow loop in a network defined by the intersections of straight-line randomly oriented source-destination paths.

c) To introduce the new concept almost-directed loop of flow in networks and formulate its basic properties.

d) To demonstrate that for flow networks (transportation networks, manufacturing networks, electrical networks and computer networks), directed and almost-directed flow loops are always associated with losses and their removal is highly beneficial.

e) To propose an efficient algorithm for identifying and removing directed loops of flow in networks with complex topology.

II. Removal Of Directed And Almost-Directed Loops Of Flow From Networks.

Denote the actual forward flows along the edges of a directed loop by Δ_{1f} , Δ_{2f} ,..., Δ_{nf} . These are all positive quantities and let the smallest among them be $\Delta_{\min} = \min\{\Delta_{1f},...,\Delta_{n-1f},\Delta_{nf}\}$. The amount $\Delta_{\min}$ will be referred to as ‘bottleneck residual capacity’.

The directed flow loop can always be 'drained' by decreasing the flow along its edges by amount $\Delta \le \Delta_{\min}$. The result is a network which is characterised by smaller losses. A 'removal' of a directed flow loop involves determining its bottleneck residual capacity $\Delta_{\min}$ and draining the loop with the amount $\Delta_{\min}$. As a result, at least one of the edges will become empty and the directed flow loop will be 'broken'.

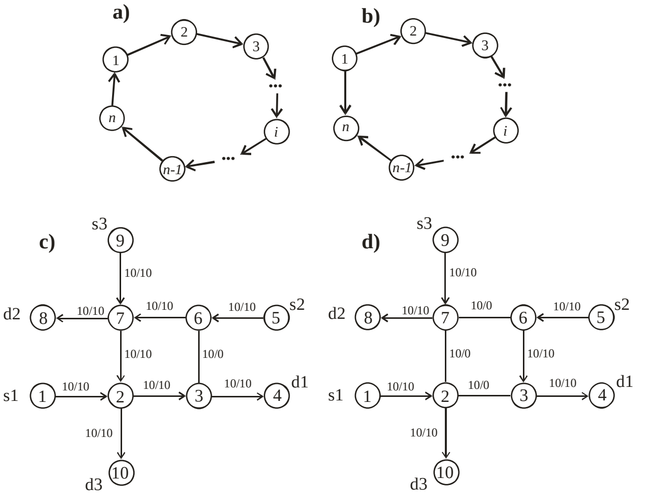


Fig. 3. a) A directed flow loop; b) An almost-directed flow loop; c,d) Removal of an almost-directed flow loop.

Suppose that there is at least one edge in the loop associated with non-zero transportation cost. The removal of flow along a directed flow loop leads to a new feasible flow and does not decrease the overall throughput flow to the destinations. In the process of flow loop removal, all edge flows have been decreased and no edge flow has been increased. Because there is at least one edge with nonzero transportation cost, the cost of transportation after the removal of the loop decreases. Thus, removing (draining) the directed flow loop (2, 3, 8, 2) carrying 100 unit of flow in the network from Fig.2, leads to a new feasible flow associated with reduced losses. The removal of the directed flow loop does not affect the throughput flow from sources to destinations.

In short, the removal of a directed flow loop always results in decreasing the losses in the network.

An almost-directed flow loop is a sequence of n sections ($n \ge 3$) in which the flow in $n-1$ edges points along the direction of traversing of the loop and the last (the closing n-th section) edge is augmentable in a direction opposite to the direction of traversing. This means that the flow in the closing edge can be increased in the direction opposite to the direction of the flow in the rest of the edges. As a result, the closing edge should not be fully saturated with flow in the direction opposite to the direction of the flow in the rest of the edges, because no flow augmentation will be possible for the closing edge.

Denote the actual flows in the forward edges by Δ_{1f}, Δ_{2f} ,..., Δ_{n-1f} and the residual space (not occupied with flow) in the closing edge by Δ_{nb}. These are all positive quantities and let the smallest among them be $\Delta_{\min} = \min\{\Delta_{1f},...,\Delta_{n-1f},\Delta_{nb}\}$. Again, the amount $\Delta_{\min}$ will be referred to as 'bottleneck residual capacity'.

The almost-directed flow loop can always be drained by decreasing the flow along the edges with forward flow by amount $\Delta \le \Delta_{\min}$ and increasing the flow with the same amount Δ along the closing edge. The draining operation does not violate the flow conservation at each node and the capacity constraints at the edges and leads to a new feasible flow.

Similar to the directed flow loops, the almost-directed flow loops are also associated with losses and their removal is highly beneficial. A 'removal' of an almost-directed flow loop means determining its bottleneck residual capacity $\Delta_{\min}$ and draining the loop with the amount $\Delta_{\min}$. As a result, either one or more of the edges with forward flow will become empty or the closing edge of the loop will become fully saturated with flow. As a result, the almost-directed flow loop will be broken.

If the cost of transportation per unit distance does not vary on the different edges, draining of an almost-closed loop always results in a reduction of the losses.

This point has been illustrated in Fig.3c with the almost closed flow loop (6,7,2,3) carrying 10 units of flow. The first label on the edges denotes the edge capacity and the second label – the actual flow through the edge. Flow of magnitude 10 units can be removed from the edges with forward flow and the flow along the closing edge (3,6) can be increased by 10 units. The result is the network in Fig.3d which is characterised by smaller losses.

Suppose that the cost of transportation per unit distance does not vary on different edges. The following theorem can then be stated.

Theorem 1. The removal of an almost-directed flow loop results in decreasing the losses in the network.

Proof. Denote the cost of transportation per unit length by c. Consider node 1 and node n. The edges (1,2), (2,3),...,(n-1,n) form a polygonal path between nodes (points) 1 and n. Denote the length of these sections (edges) by l_{12}, $l_{2,3}$,..., $l_{n-1,n}$. Denote the length of the closing section (edge) by $l_{n,1}$. The length of a polygonal path between two points however, is greater than the length of the distance between the two points. Therefore,

$$\sum_{i=1}^{n-1} l_{i,i+1} > l_{n,1} \qquad (1)$$

holds for a polygonal path which does not degenerate into a straight line. Because the cost of transportation per unit length is the same, removing the bottleneck flow $\Delta_{\min}$ from the almost-directed loop will result in a reduction of the transportation cost by $-c\,\Delta_{\min}\sum_{i=1}^{n-1} l_{i,i+1}$ along edges (1,2), (2,3),...,(*n*-1,*n*) and an increase of the transportation cost by $c\Delta_{\min} l_{n,1}$ along edge (*n*,1). Considering inequality (1), the inequality

$$-c\,\Delta_{\min}\sum_{i=1}^{n-1} l_{i,i+1} + c\Delta_{\min} l_{n,1} = c\Delta_{\min}(l_{n,1} - \sum_{i=1}^{n-1} l_{i,i+1}) < 0$$

is then valid, which means that removing the almost-directed loop will result in a net decrease of the cost of transportation and therefore in reduction of the losses.□

The next theorem permits the removal of directed flow loops and almost-directed flow loops to proceed in two stages. During the first stage only directed flow loops are removed while during the second stage, only almost-directed flow loops are removed.

Theorem 2. *If there are no directed flow loops in the network, the removal of an almost-directed flow loop cannot possibly create a directed flow loop.*

Proof. Suppose that the removal of the almost-directed loop (1,...,k,...,n) created a directed flow loop (Fig.4). Initially, by assumption, no directed flow loops exist in the network. Because the flow through the entire (1,k,n) section has been decreased by the removal of the almost-directed loop (1,k,n,1), a new directed flow loop can only appear if the closing edge (n,1) has been initially empty and after the increase of its flow from node 1 to node n, is now part of the new directed loop (n,p,1,n), (Fig.4).

Let (1,n,p,1) be the new directed flow loop. This is however impossible because before the removal of the almost-directed loop (1,k,n,1), a concatenation of sections (n,p,1) and (1,k,n) would have created a directed flow loop. This contradicts the assumption that no directed flow loops exist initially, therefore the theorem is true.

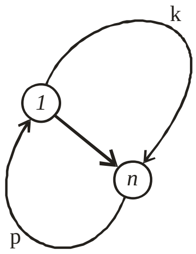


Fig. 4. In a network without directed flow loops, the removal of an almost-directed flow loop cannot possibly create a directed flow loop.

Finally, it can be shown that the process of removing almost-directed flow loops is finite for networks with integer capacities and must terminate.

Indeed, if the number of edges is m and the largest edge capacity is C, the maximum possible flow quantity contained in the network is mC. At each removal of an almost-directed flow loop, at least 1 unit of flow is removed from more than one edge and the same amount of flow is added to the single closing edge of the almost-directed loop. Consequently, the net change of the flow is negative and the amount of flow contained in the network decreases by at least 1 unit. As a result, after a finite number of steps, the process of removing the almost-directed loops will terminate.

III. Estimating The Likelihood Of A Directed Flow Loop In A Network Formed By The Intersections Of Randomly Oriented Straight-Line Source-Destination Paths

The unexpectedly high probability of existence of directed flow loops will be demonstrated by considering the general case where the source-destination paths are randomly oriented intersecting straight lines (Figure 5a).

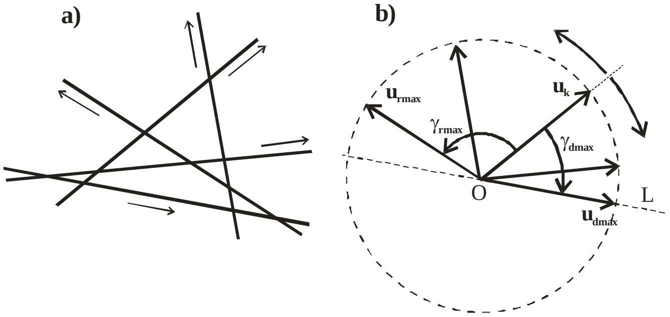


Fig. 5. a) Randomly oriented intersecting source-destination paths; b) All direction vectors of the source-destination paths can be translated to start from a common point *O*.

All source-destination paths transport the same type of interchangeable commodity (e.g. petrol) and for each source-destination path; there is a particular direction of the flow (Fig.5a).

It is assumed that there are at least three source-destination paths; there are no parallel paths and no three paths intersect into a single point. These conditions are natural and common. Indeed, for randomly oriented straight-line paths on a plane, it is very unlikely to find two parallel paths or three paths intersecting into a single point.

The likelihood that a directed flow loop will be present in the network, given that the orientation of the intersecting source-destination flows is random, will be termed 'probability of a directed flow loop for random source-destination paths'.

The existence of a directed flow loop anywhere between the points of intersection implies the existence of a triangular directed flow loop (Fig.6a). As a result, the existence of a triangular directed flow loop is a necessary condition for the existence of a flow loop. Conversely, the existence of a triangular directed flow loop is also a sufficient condition for the existence of a flow loop. Consequently, the probability of a directed flow loop for randomly oriented source-destination flows can be estimated by estimating the probability of a directed triangular flow loop – the existence of three intersection points, between which the flow travels in the direction of traversing these points (Fig.6a).

A unit vector can be assigned to each source-destination path, whose direction is the same as the direction of the flow along the path (Fig.5b). The angle α the unit vector subtends with the horizontal axis (Fig.7a) gives the orientation of the source-destination path and the direction of the flow along the path. A random orientation of a source-destination path and the direction of its flow means that the angle α the direction vector subtends with the fixed horizontal x-axis is uniformly distributed in the interval (0,2π). In other words, all possible orientations are characterised by the same probability.

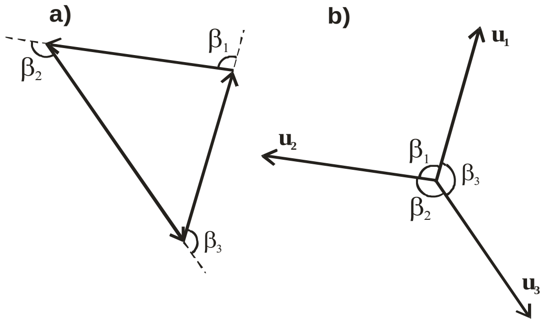


Fig. 6. a) A directed triangular flow loop; b) direction vectors of the source-destination paths forming the directed flow loop.

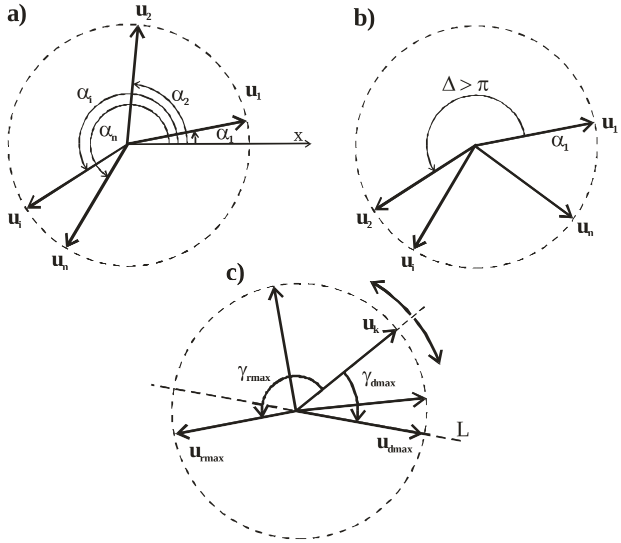


Fig. 7. a) Ordering the direction vectors, according to the angle they subtend with the horizontal axis; b) a gap of size at least π between two random direction vectors; c) If no half-plane can be selected where all direction vectors reside, there is always a possibility to select three direction vectors which do not reside in a single half-plane.

The unit vectors assigned to the source-destination paths will be referred to as 'direction vectors'. They can all be translated at the common origin *O*, as shown in Fig.5b. The following theorem then holds.

Theorem 3. The necessary and sufficient condition for a directed flow loop in a network defined by the intersections of randomly oriented straight-line source-destination paths, is the non-existence of a half-plane where all direction vectors reside.

Before proving this theorem two lemmas will be stated and proved.

Lemma 1. If any three selected direction vectors reside in a single half-plane, then all direction vectors reside in a single half-plane.

Proof. Let us select an arbitrary direction vector $\mathbf{u}_\mathbf{k}$ and introduce counterclockwise and clockwise direction with respect to this vector to mark the angular positions of the rest of the direction vectors (Fig.5b). The angles γ_{di} mark the position of the unit vectors located in a clockwise direction from the vector $\mathbf{u}_\mathbf{k}$ up to an angle equal to π. The angles γ_{ri} mark the positions of the unit vectors located in a counter-clockwise direction from the unit vector $\mathbf{u}_\mathbf{k}$ up to an angle equal to π.

Let $\gamma_{d\max}$ be the angle corresponding to the most extreme direction vector $\mathbf{u}_\mathbf{dmax}$ in a clockwise direction and $\gamma_{r\max}$ be the angle corresponding to the most extreme direction vector $\mathbf{u}_\mathbf{rmax}$ in a counterclockwise direction. By assumption, any three direction vectors reside in a single half-plane, therefore the three direction vectors $\mathbf{u}_\mathbf{k}$, $\mathbf{u}_\mathbf{dmax}$ and $\mathbf{u}_\mathbf{rmax}$ also reside in a single half-plane. Consequently, $\gamma_{d\max} + \gamma_{r\max} < \pi$ and the three vectors $\mathbf{u}_\mathbf{k}$, $\mathbf{u}_\mathbf{dmax}$ and $\mathbf{u}_\mathbf{rmax}$ reside in the half-plane defined by the line *L* (oriented along the unit vector $\mathbf{u}_\mathbf{dmax}$) and the unit vector $\mathbf{u}_\mathbf{k}$ (Fig.5b). Because the rest of the unit vectors reside either within the angle $\gamma_{d\max}$ or within the angle $\gamma_{r\max}$ ($\gamma_{d\max}$ and $\gamma_{r\max}$ are the extreme angles corresponding to the direction vectors), all of the direction vectors must also reside in the half-plane defined by the line *L* and the unit vector $\mathbf{u}_\mathbf{k}$. □

Lemma 2. *If no half-plane can be selected where all direction vectors reside, there is always a possibility to select three direction vectors which do not reside in a single half-plane.*

Proof. Similar to the previous proof, an arbitrary direction unit vector $\mathbf{u}_\mathbf{k}$ is selected and counterclockwise and clockwise direction with respect to this vector is introduced to mark the angular positions of the rest of the direction vectors (Fig.7c). Again, $\gamma_{d\max}$ marks the most extreme direction unit vector $\mathbf{u}_\mathbf{dmax}$ in clockwise direction up to an angle equal to π and $\gamma_{r\max}$ marks the most extreme unit vector $\mathbf{u}_\mathbf{rmax}$ in counterclockwise direction up to an angle equal to π. The three vectors $\mathbf{u}_\mathbf{k}$, $\mathbf{u}_\mathbf{dmax}$ and $\mathbf{u}_\mathbf{rmax}$ do not reside in a single half-plane (Fig.7c).

Indeed, suppose that the three vectors $\mathbf{u}_\mathbf{k}$, $\mathbf{u}_\mathbf{dmax}$ and $\mathbf{u}_\mathbf{rmax}$ reside in a single half plane. As a result, $\gamma_{d\max} + \gamma_{r\max} < \pi$ and the three vectors $\mathbf{u}_\mathbf{k}$, $\mathbf{u}_\mathbf{dmax}$ and $\mathbf{u}_\mathbf{rmax}$ should reside in the half-plane defined by the line *L* (oriented along the direction vector **u_{dmax}**) and vector **u_k**. Because $\gamma_{d\max}$ and $\gamma_{r\max}$ are the extreme angles corresponding to the separate direction vectors, the rest of the unit vectors reside either within the angle $\gamma_{d\max}$ or within the angle $\gamma_{r\max}$. As a result, all of the direction vectors must also reside in the half-plane defined by the line *L* and the unit vector $\mathbf{u_k}$. This is however impossible because, according to our assumption, no half-plane can be selected where all direction vectors reside. This contradiction shows that the selected direction vectors $\mathbf{u_k}$, $\mathbf{u_{dmax}}$ and $\mathbf{u_{rmax}}$ do not reside in a single half plane.□

Now, Theorem 3 can be proved.

Proof of Theorem 3. First note, that the existence of a directed flow loop implies the existence of at least one triangular directed flow loop (Fig.6a) because no two source-destination paths are parallel. Suppose that the source-destination paths with direction unit vectors $\mathbf{u_1}$, $\mathbf{u_2}$ and $\mathbf{u_3}$, (Fig.6b) define a triangular directed flow loop and the angles between the source-destinations paths from Fig.6a are β_1, β_2 and β_3. Because the intersecting paths form a triangle, the sum of the angles β_1, β_2 and β_3 is always exactly equal to 2π (Fig.6b).

$$\beta_1 + \beta_2 + \beta_3 = 2\pi \tag{1}$$

In addition, for each angle β_i , the conditions

$$0 < \beta_i < \pi \text{ , } i = 1,2,3 \tag{2}$$

Are always fulfilled. Suppose that there is a single half plane where the direction vectors of all source-destination paths reside. In this case, the direction vectors $\mathbf{u_1}$, $\mathbf{u_2}$ and $\mathbf{u_3}$ of the paths forming the directed triangular loop will also reside in the same half-plane. However, this is impossible because in this case, the conditions (1)-(2) will be violated.

Now, suppose that there is no half-plane where all of the direction vectors $\mathbf{u_1}$, $\mathbf{u_2}$,..., $\mathbf{u_n}$ reside. According to Lemma2, in this case, we can always select three vectors u1, $\mathbf{u_2}$ and $\mathbf{u_3}$ which do not reside in the same half-plane. For these three vectors, conditions (1) and (2) will be fulfilled. Because, by assumption, no three source-destination pairs intersect in a single point, the source-direction paths which correspond to the selected direction vectors $\mathbf{u_1}$, $\mathbf{u_2}$ and $\mathbf{u_3}$ will form a triangular directed flow loop.□

Theorem 3 serves as a basis for calculating the probability of a directed flow loop. This probability can be determined by determining first the probability of the complementary event that no directed flow loop exists.

To calculate this probability, the random direction vectors are ordered in ascending order, according to the magnitude of the angle they subtend with the horizontal *x*-axis (Fig.7a).

The probability that there will be no directed flow loop is equal to the probability that all random direction vectors will lie in a single half-plane. All random direction vectors will lie in a single half-plane if and only if a gap of minimum length $\Delta = \pi$ exists between two random direction vectors (Fig.7b). There can be no more than a single gap of size $\Delta = \pi$, therefore, the random events corresponding to a gap between the first and the second direction vector, between the second and the third direction vector, etc., are mutually exclusive events. As a result, a single gap of minimum size $\Delta = \pi$ may be located in *n* distinct, mutually exclusive ways between the direction vectors.

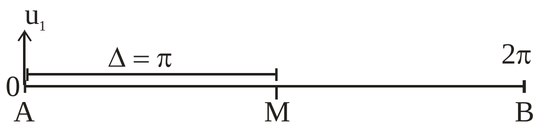


Fig. 8. A gap of length Δ can be located in *n* distinct ways between the direction vectors.

Let the circumference of the direction vectors circle be represented by the segment with length 2π (Fig.8). A gap of length $\Delta = \pi$ between direction vectors u_1 and u_2, can only occur if the rest of the *n*-1 random locations fall in the segment MB and none of them falls in the segment AM (Fig.8). Because the probability that a random direction vector location will 'fall' on the segment MB is $\pi/(2\pi) = 1/2$, the probability that *n*-1 random vector locations will fall in the segment MB is $1/2^{n-1}$.

Similarly, the probability of a gap between the second and the third direction vector is also $1/2^{n-1}$. As a result, the probability p of a gap of minimum size $\Delta = \pi$, between two direction vectors is a sum of the probabilities of these mutually exclusive events and becomes

$$p = \sum_{k=1}^{n} 1/2^{n-1} = n/2^{n-1} \tag{3}$$

Which is also the probability that no directed flow loop will exist. Because there can be either a directed flow loop or no directed flow loop, the probability of a directed flow loop is:

$$\Pr(\textit{directed flow loop}) = 1 - n/2^{n-1} \tag{4}$$

The probability of existence of a directed flow loop from equation (4) has been plotted in Fig.9.

As can be verified, with increasing the number of intersecting random source-destination paths, the probability of a directed flow loop increases significantly. For five intersecting flow paths, the probability of a directed flow loop is already 69%.

Note from equation (4) that no matter how large the number n of intersecting source-destination paths is, the probability of a directed flow loop is always smaller than 1.

This means that for any possible number and for any possible orientation of straight-line flow paths on a plane, it is always possible to choose the directions of the flows along the paths in such a way, that no parasitic flow loops appear between the points of intersection.

The results from equation (4) have been confirmed by a Monte Carlo simulation.

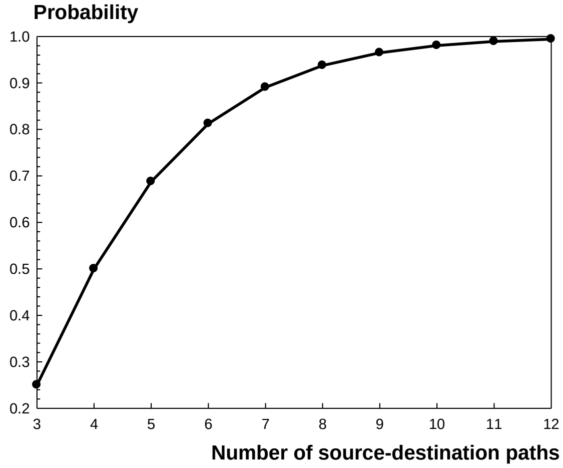


Fig. 9. Probability of a directed flow loop as a function of the number of intersecting source-destination paths.

IV. Discovering And Removing Directed And Almost –Directed Loops Of Flow In Networks

The algorithm for discovering and removing a directed flow loop is based on calling a depth-first-search (dfs) procedure from an initial node and subsequently calling the dfs-procedure from the descendents of this node, etc., until an already traversed node is discovered again. Initially, all nodes are marked as 'not visited' (coloured' white). As the nodes are visited by the dfs-procedure, they are marked as "visited" (coloured gray, Fig.10). We must point out that the dfs-procedure does not scan all successors of the current node. It scans all eligible successors only. A successor node *i* of the current node 'r_node' is eligible if: (i) there is an edge directed from node r_node to node *i* and the edge (r_node,i) carries nonzero flow. If edge (r_node, *i*) is empty, the successor *i is not considered by the dfs-procedure*. Suppose that the call of the dfs- procedure from node 'r_node1' does not discover any already traversed node (Fig.10a). In this case, after the return from the dfs-call, the node r_node1 is marked as 'completed' (coloured black) (the entry of the r_node1 in the array cmpl[] is set to '1', cmpl[r_node1]=1). Under these conditions, the following theorem holds.

Theorem 4. *In a network with feasible flow, a directed flow loop is present only if during a recursive call of the depth-first search procedure, an already traversed node has been discovered again and it has not been marked as 'completed'.*

Proof. Suppose that the node '*e*', which has been visited again, has been marked as 'completed' (Fig.10a). Because the node has been marked as 'completed' (cmpl[*e*]=1), an earlier depth-first search must have started from this node and no directed flow loop must have been discovered starting with node '*e*'. Therefore, node '*e*' discovered twice and marked as 'completed', cannot possibly belong to a directed flow loop.

Now suppose that the dfs-procedure has been called and during the subsequent calls of the dfs-procedure from subsequent descendent nodes, an already traversed node 'r_node' not marked as 'completed' (cmpl[r_node]=0), has been visited again (Fig.10b). Because node 'r_node' has been visited for a second time and it has not been marked as 'completed', no return from the earlier dfs-call initiated from this node has occurred (otherwise the node would have been marked as 'completed'). As a result, the 'r_node' has been visited again, after traversing a chain of descendent nodes starting from node 'r_node' and ending at r_node (Fig.10b). This essentially means that a directed flow loop has been discovered. □

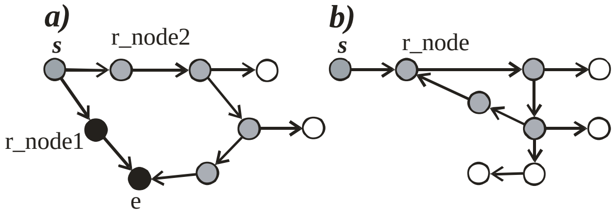


Fig. 10. Traversing the nodes of the network by recursive calls to the depth-first-search procedure.

Here is the algorithm in pseudo-code:

Direct all edges of the network to match the directions of the edge flows.
//As a result, the network is transformed from a network with undirected edges to a network with directed edges.

```
procedure retrieve_directed_flow_loop(cur_node)
{// retrieves and eliminates the identified directed
   loop of flow}

procedure dfs(r_node)
{
  marked[r_node] = 1;
  for i= 1 to all eligible successors of r_node do
  {
    cur_node = current eligible successor;
    if (marked[cur_node] = 0) then {
                    pred[cur_node] = r_node;
                    dfs(cur_node);
                    }
    else { if (cmpl[cur_node] = 0) then
           {
             pred[cur_node] = r_node;
             retrieve_directed_flow_loop(cur_node);
             break;
           }
         }
  }
  cmpl[r_node] = 1;
}
```

Statements before the call of the dfs-procedure:
for i=1 **to** n **do** {marked[i] = 0; cmpl[i] = 0; pred[i] = 0;}
dfs(1).

A directed flow loop can only be discovered if both conditions are fulfilled: (i) a node cur_node, already marked as 'visited' has been encountered during the search and (ii) the call dfs(cur_node) is still active, in other words, its activation record is still in the stack. After the end of the dfs (cur_node) call, node r_node is marked as 'completed' by the statement 'cmpl[r_node] = 1'. This is why, only when both marked[cur_node] = 1 and cmpl[cur_node] = 0 are encountered during the search, a directed loop of nonzero flow has been discovered. The directed loop of flow is subsequently retrieved and eliminated by the procedure ***retrieve_directed_flow_loop***(cur_node). The array pred[] records the predecessors of the visited nodes and helps retrieve the identified directed flow loop. The procedure ***retrieve_directed_flow_loop***(cur_node) retrieves the discovered loop of flow by starting with the statement 'k = cur_node' followed by a loop, where the statement k = pred[k] is repeatedly executed and followed by a check whether an already encountered node has been encountered again and whether node *k* is a descendent of the 'cur_node'. These checks are used for identifying the node which closes the identified directed loop carrying nonzero flow. After discovering the directed flow loop, the procedure determines the edge from the loop carrying the smallest amount of flow and subtracts this flow from the flows of all edges belonging to the loop. As a result, at least one edge in the directed loop becomes empty. Once an edge becomes empty, it remains empty until the end of the procedure for removing directed flow loops. This flow loop will not be discovered again during subsequent searches.

The proposed algorithm has been tested on a benchmark network which has the shape of a lattice (Fig.11). The lattice-type network has been selected because it is a natural network, often encountered in real applications. Because the lattice network is easily scalable, it provides an opportunity to isolate the impact of the size of the network on the algorithm's performance. To increase the number of loops, alternating directions of the flows from the sources s_i to the destinations d_i has been selected (Fig.11).

The same flow of 10 units per unit time has been assumed along each source-destination path.

Six different sizes of lattice-type network have been constructed and the network loops have been removed by the proposed algorithm, implemented in *C* and run on a computer with a processor *Intel(R) Core(TM) 2 Duo CPU T9900 @ 3.06 GHz*.

According to the number of intersecting horizontal and vertical paths, the following network sizes were tested: 2x2, 3x3, 4x4, 5x5, 6x6 and 7x7. The number of nodes corresponding to the 6 test-networks was: 12, 22, 32, 45, 60 and 77, correspondingly.

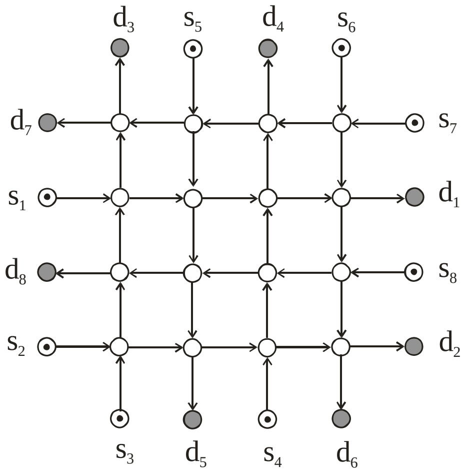


Fig. 11. Lattice networks used for testing the proposed algorithm.

The running time of the algorithm versus the size (the number of nodes) of the lattice network are shown in Fig.12. As can be seen, the running time of the algorithm is approximately proportional to the size of the lattice network.

It needs to be pointed out that identifying all directed cycles in the network, before removing the bottleneck flow from any of them, is not a feasible approach. To show why this is the case, consider a complete network where any two nodes (i,j) are connected with directed edges (Fig.13 shows a complete network with 4 nodes). The number of directed cycles in this network is equal to the number all possible subsets of 2 nodes, 3 nodes,...,n nodes. Consequently, the number of directed cycles is $2^N - N - 1$ and determining all possible cycles is a task of exponential complexity, a task which is practically impossible even for not very large *n*. In the proposed approach, identifying a directed loop of flow with the dfs-procedure and subtracting the bottleneck flow from the edges, has a worst-case complexity O(*m*), where *m* is the number of edges in the network. After each flow subtraction, at least a single edge remains empty. Therefore, after at most *m* steps, all directed loops of flow will be discovered and removed. As a result, the procedure for removing all directed loops of flow in a network has a worst-case running time O(m^2).

Finally, it can be shown [6] that maximising the throughput flow at a minimum cost, leaves no directed loops of flow in the network. The worst-case running time of maximising the throughput flow at a minimum cost however is significantly larger than the worst-case running time of the described procedure.

Similar to the case related to directed flow loops, the probability of existence of almost-directed loops of flow in networks can be estimated and an algorithm related to discovering and removing almost-closed flow loops can be developed. The developments related to almost-directed flow loops will be published elsewhere.

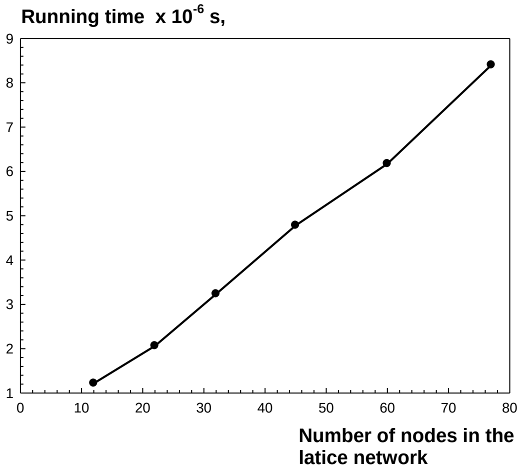


Fig. 12. Performance of the proposed algorithm for a different size of the lattice network.

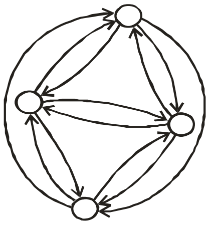


Fig. 13. Complete network with four nodes

V. Conclusions

1) Directed loops of flow can appear in networks with interchangeable commodity even if no transported commodity physically travels along a closed contour.

2) The necessary and sufficient condition for a directed flow loop in a network defined by randomly oriented straight-line flow paths, is the non-existence of a half-plane where all direction vectors reside.

3) A closed-form expression has been obtained for the probability of a directed flow loop for intersecting, randomly oriented flow paths, in a plane. Even for a relatively small number of intersecting flow paths, the probability of a directed flow loop is very large.

4) A theorem has been stated and proved regarding the existence of a directed flow loop in a network with feasible flows. On the basis of this theorem, a simple and efficient recursive algorithm has been proposed for discovering and removing directed loops of flow in networks. The algorithm discovers and removes a directed flow loop in linear time in the size of the network.

5) A new concept referred to as 'almost-directed flow loop' has been introduced and its basic properties formulated. It has been shown that the removal of an almost-directed flow loop results in decreasing the losses in the network.

6) It is shown, that the process of removal of almost-directed flow loops terminates after a finite number of steps. Furthermore, if no directed flow loops exist in the network, the removal of an almost-directed flow loop cannot create a directed flow loop.

7) The shortest-path strategy for optimising the throughput flow between sources and destinations does not guarantee that there will be no undesirable directed loops of flow in the optimised networks.

8) The directed and almost-directed flow loops in real networks are associated with wastage of energy and resources and increased levels of congestion. In real networks (transportation networks, manufacturing networks, electrical networks, computer networks, etc.), which include many intersecting flow paths, the existence of directed and almost-directed flow loops and the associated wastage of energy for maintaining these loops is practically inevitable. Consequently, optimising real networks by removing directed and almost-directed flow loops has the potential to save a significant amount of wasted resources for the world economy.

References

[1] U. Hengartner, S. Moon, R. Mortier, and C. Diot. Detection and analysis of routing loops in packet traces. *Sprint ATL Technical Report TR02-ATL-051001*, May 2002.

[2] V. Paxson. End-to-end routing behavior in the Internet. *IEEE/ACM Transactions on Networking*, vol. 5(5), pp.610–615, 1997.

[3] Tanenbaum A.S., Computer networks, 4th ed., Pearson Education International, 2003.

[4] Sharir M., A strong-connectivit;y algorithm and its applications in data flow analysis, Computers and Mathematics with Applications, vol.7, pp.67-72, 1981.

[5] Todinov M.T., The dual network theorem for static flow networks and its application for maximising the throughput Flow, Artificial Intelligence Research, vol.2 (1), pp.81-106, 2013.

[6] Todinov M.T., Flow Networks, Elsevier, 2013.

[7] Ahuja R.K., T.L.Magnanti, J.B.Orlin, Network flows: Theory, Algorithms and Applications, Prentice Hall, 1993.

[8] Asano T., Y.Asano, Journal of the Operations Research Society of Japan, vol.43 (1), 2000.

[9] Goodrich M.T. and R.Tamassia, Algorithm design, John Wiley & Sons, 2002.

[10] Hu T.C., Integer programming and network flows, Addison-Wesley Publ.Company, Reading, Mass., 1969.

[11] Cormen T.H., T.C.E.Leiserson, R.L.Rivest, and C.Stein, Introduction to Algorithms, 2nd ed., MIT Press and McGraw-Hill, 2001.

[12] Tarjan R.E., Data structures and network algorithms, SIAM, Philadelphia, 1983.

[13] Kleinberg J. and E.Tardos, Algorithm design, Addison Wesley, 2006.

[14] Lawler E., Combinatorial Optimisation, Dover publications, Inc, New York, 1976.

[15] Goldberg A.V., E.Tardos and R.E.Tarjan, Network flow algorithms, in Algorithms and Combinatorics, v.9. Paths, Rows and VLSI-Layout, Sprmger-Verlag Berlin, Heidelberg, 1990.

[16] Papadimitriou C.H., K.Steiglitz, Combinatorial Optimisation: Algorithms and complexity, Dover publications, Inc., New York, 1998.

[17] Dinic E.A., "Algorithm for solution of a problem of maximum flow in a network with power estimation", Soviet Math. Doklady, vol. 11(8), pp.1277-1280, 1970.

[18] Edmonds J. and R.M.Karp, Theoretical improvements in algorithmic efficiency for network flow problems, Journal of the ACM, vol.19(2), pp.248-264, 1972.

[19] Elias P., A.Feinstein and C.E.Shannon, Note on maximum flow through a network, IRE Transactions on Information Theory, IT2, pp.117-119, 1956.

[20] Ford L.R. and D.R. Fulkerson, Maximal flow through a network, Canadian Journal of Mathematics, vol.**8**(5), pp.399-404, 1956.

[21] Karzanov A., Determining the maximal flow in a network by the method of preflows., Sov.Math.Dokl., 1974.

[22] Goldberg A.V. and R. E. Tarjan, A new approach to the maximum flow problem, Journal of ACM, vol.**35**, pp.921-940, 1988.

[23] Sleator D.D. and R.E.Tarjan. An O(*nm* log *n*) algorithm for maximum network flow. Technical report STAN-CS-80-831, Department of Computer Science, Stanford University, Stanford, CA, 1980.

Workshop Session Recordings on Green Volunteering Activities of Students in a Disadvantaged Area According to the Good-Hearted Vocation Teacher to Support Itinerant Junk Buyers

Kuntida Thamwipat
Faculty of Industrial Education and Technology
King Mongkut's University of Technology Thonburi
Bangkok, Thailand

Thanakarn Kumphai
Faculty of Industrial Education and Technology
King Mongkut's University of Technology Thonburi
Bangkok, Thailand

Abstract—**This project was aimed to provide workshop session recordings on green volunteering activities of students in one disadvantaged area under the bridge of zone 1, Pracha-Utit Road 76, Toong-kru District, Bangkok where the majority worked as itinerant junk buyers. Therefore, the students held workshop sessions with the aim to provide training on how to repair electrical appliances and engines so that the community members could use this knowledge to increase the value of the unwanted electrical appliances they bought. The project also discussed the risk and danger of certain junk product which might be mixed with rubbish and taught how to classify recyclable products to increase the value of the junk. This project was the first of its own and it was done as green volunteering activities of students. The research team has provided 182 families from the community under the bridge of zone 1 with a number of workshop sessions. The sampling group was chosen out of those who attended at least three times and there were 20 persons. The research results showed that the sampling group achieved high level of knowledge (100.0%). They could fix fans as well as repair and maintain engines. They could classify junk. They expressed high level of satisfaction towards the workshop sessions (mean score of 4.18 with S.D. of 0.27). When the assessment was conducted as regards the operation and the recordings on green volunteering activities of 13 students, it was at the highest level (mean score of 4.68 with S.D. of 0.42). This workshop project was the first runner-up of the national SCB Challenge 2012 Community Project as organized by Siam Commercial Bank PLC.**

Keywords—Recordings; Workshop; Student Activities; Green Volunteering; Disadvantaged Community

I. BACKGROUND

The communities under the bridge in many parts of Bangkok are places where many homeless lived. Bangkok Metropolitan Administration and National Housing Agency have collaborated to provide over 700 families of these people with 3 plots of land which are not far from their previous places. These are the community under the bridge at Pracha-Utit 76, Toong-kru District (Zone 1), the community under the bridge at Poonsarp, Saimai District (Zone 2) and the community under the bridge at Onnuch, Prawate District (Zone 3). The community under the bridge of zone 1 is located at Pracha-Utit Road Soi 76, Toong-kru District, Bangkok which is about 10 kilometres away from King Mongkut's University of Technology Thonburi. This 13-rai* (*1 rai is equal to 1,600 square metres) plot of land houses 182 families at the moment. There are public areas such as sports field, playground, and pre-school development centre to hold meetings and activities among family members. The majority of people or over 70% of them work as itinerant junk buyers, in other words, they buy and collect unwanted or faulty electrical appliances, litter, empty plastic bottles, paper and the like and then they classify and sell them later. The majority of people are poor and their educational level was not high. They use saleng or three-wheeled pedal cart as their vehicle and as such, their community is sometimes called "Saleng Community" which is one of many disadvantaged communities in Thailand.

In 2011, King Mongkut's University of Technology Thonburi (KMUTT) conducted the research study entitled "Community Research Project to Reduce and Solve the Social Inequality in Bangkok: A Case Study of Community under the Bridge of Zone 1, Toong-kru District, Bangkok" [1] with National Institute of Development Administration and Bangkok Metropolitan Administration in order to examine and analyse the current situations and requirements of the community. The results from this research included 15 developmental policy plans which had been amended by the community commission and the community people. The main aim is to develop the community continuously. The working group in this project consisted of 13 second-year and third-year undergraduate students from Faculty of Industrial Education and Technology, King Mongkut's University of Technology Thonburi under supervision by their advisors to depict the problems of the disadvantaged community. Therefore, the working group proposed to hold workshop sessions about green volunteering activities entitled "Good-Hearted Vocation Teacher to Support Itinerant Junk Buyers" to offer training sessions on repairing electrical appliances and engines so that the community members could apply this knowledge to their profession, namely buying unwanted or faulty products. They could fix faulty products to sell with more value and maintain their saleng or pedal cart to reduce the maintenance cost.

Besides, the workshop sessions were also aimed at developing the bodies of knowledge and the potentials of the working group. Since the students were studying electrical education and mechanical education to become vocation teachers in the future, they could develop their ability to teach their knowledge both in theory and in practice through these workshop sessions [2]. Apart from the knowledge about repairing electrical appliances and maintaining engines, the project also discussed the risk and danger of certain junk product and each type of rubbish [3] as well as how to give first aid. The working group hopes that the knowledge about repairing appliances and maintaining engines as well as understanding about safety and rubbish classification will meet the needs of the community so that their life conditions are of better quality.

II. Research Objectives

A. To hold workshop sessions about how to repair electrical appliances, how to maintain engines and how to classify rubbish for the people in the community under the bridge of zone 1.

B. To measure the knowledge about how to repair electrical appliances, how to maintain engines and how to classify rubbish of the people in the community under the bridge of zone 1.

C. To examine the satisfaction towards the workshop sessions about how to repair electrical appliances, how to maintain engines and how to classify rubbish as expressed by the people in the community under the bridge of zone 1.

D. To assess the results from the operation and the recordings about green volunteering activities of students who participated in the project.

III. Research Scope

The data in this research were collected in the second term of the academic year 2011 between January and February 2011 in WatPuttabucha Market and nearby communities only.

A. Skills in repairing electrical appliances: There were 3 parts in this training session:

1) Basic understanding about electrical appliances and repair kit.

2) How to repair household appliances, including rice cooker, kettle, iron and fan..

3) How to install electric wiring.

B. Skills in repairing and maintaining engines: There were 2 parts in this training session:

1) How to notice the faults and how to fix engines, including basic understanding about motorcycle electric system, water soaked engine, and engine basics.

2) How to replace motor devices such as replacing oil, mending punctures, and replacing tyres.

C. Understanding about various types of rubbish and how to classify them

1) Understanding about each type of rubbish.

2) Principles and technique in classifying rubbish.

3) Danger and risk from rubbish.

4) First aid for those in danger of rubbish.

To run the project and each workshop session, the working group of students and presenters would provide the community members with knowledge and training under the supervision of advisors as shown in Table 1.

TABLE I. Shows the Roles and Responsibilities in Each Activity

Activity/Phase	The working group of 13 students	Advisory Board	Benefits to the Community
Preparation and Data Collection	Data and contents were collected for the project/ workshop	The appropriateness of the contents were considered	-
Campaigning for the project at the site	Media were created and the campaign was done at the site	Field trip with the students to inform the community commission of the details	The community members got information and prepared themselves for the workshop
Workshop session 1: How to repair electrical appliances	Electrical students ran the workshop	Field trip and guidance to students	The community members gained knowledge and hands-on experience with electrical appliances
Workshop session 2: How to repair and maintain engines	Mechanical students ran the workshop	Field trip and guidance to students	The community members gained knowledge and hands-on experience with engines
Workshop session 3: Types of rubbish and how to classify them	Guest speakers to run the workshop while students supported them	Field trip and guidance to students	The community members applied this knowledge to the safety issues in their profession
Follow-up and Assessment	Interviews and data collection through questionnaire	Field trip and interviews with the community commission	-
Report on the Operation	Results from the operation were gathered and written up in the final report	Report approval	-

IV. Population and the Sampling Group

The population in this study were 182 families living in the community under the bridge of zone 1, Pracha-Utit 76 and working as itinerant junk buyers. The sampling group was chosen using purposive sampling method out of those who attended at least 3 workshop sessions. There were 20 persons in total.

V. Tools For Data Collection

A. *The observation form to measure the level of understanding after the workshop*

B. *The questionnaire with Likert's 5-rating scale to measure the satisfaction towards the workshop*

C. *The self-assessment form for the working group*

VI. Statistical Methods Used

Percentage, Mean and Standard Deviation.

VII. Research Results

A. *Workshop sessions on repairing electrical appliances, fixing engines and how to classify rubbish*

1) Unwanted? Give us! ←Preparation→ Community (Field Trip, Survey of what's needed, Meetings, Review, Safety Training).

2) Public relations.

3) Teaching Preparation and Plan.

4) Training→ Sessions 1-6: Repairing Electrical Appliances, Sessions 7-9: Repairing Engines, Session 10: Classifying Rubbish and Safety, Session 11: Painting the field.

5) Trainees need to repair electrical appliances within the time limit ←Assessment with Electrical Appliances.

6) Summary→ Number of attendants, Community technicians, Satisfaction.

7) Follow-up→ Statistical data.

The project began the field trip and ran the operation from October 2012 to January 2013.

B. *The results about the level of understanding for 3 workshop sessions*

The level of understanding was assessed individually during the time set for the test in which the participant had to repair or fix the faulty parts in the device. The assessment was done in the following details: how to use devices, how to troubleshoot the problems, how to choose or replace the spare parts, and how to mend it according to the training session. The level of understanding could be discussed as follows:

a) Repairing Electrical Appliances

TABLE II. Shows the Number of People and the Level of Understanding About How To Repair Electrical Appliances

Topic/ No. of persons	Maintenance	Dismantling and basic troubleshooting	Analysis and repair	Note
Fan	3	4	20	100 %
Iron	5	7	11	
Rice cooker	5	7	11	
Kettle	5	7	11	
TV	1	15	5	To repair TV needs a lot of details and more time

b) Repairing and Maintaining Engines

TABLE III. Shows the Number of People and the Level of Understanding About How To Repair and Maintain Engines

Topic/ No. of persons	Maintenance and basic check-up	Analysis and replacement	Note
Replacing oil, Cleaning carburetor	5	20	100 % (Everybody could do it because they were familiar with saleng)
Mending puncture, Fixing chains	5	20	100 % (Everybody could do it because they were familiar with saleng)

c) Classifying Rubbish and Safety when Handling Rubbish and Junk

The trainees were familiar with the classification of rubbish. When they took a game test, they could win it. As for the topic of occupational health, it was still new to the community. After the training and the test on their understanding as well as the game activities and interviews, everybody, both adults and children, gained the highest level of understanding.

C. *Results about the Satisfaction towards the Green Volunteering Activities Organized by Students*

The results about the satisfaction towards the green volunteering activities organized by students by the sampling group of 20 persons from different age groups and genders in the community were as follows:

TABLE IV. Shows the Satisfaction towards the Green Volunteering Activities Organized by Students

Item	Satisfaction level		
	Mean	S.D.	Meaning
Speakers			
1. Clarity in knowledge transfer	4.50	0.50	High
2. Ability to explain the contents	4.25	0.78	High
3. Connection of contents and workshop	4.35	0.74	High
4. Comprehensiveness	4.40	0.50	High
5. Time efficiency	4.05	0.82	High
6. Questions and feedbacks to trainees	4.35	0.67	High
Average	4.32	0.62	High
Location/ Duration/ Catering			
1. Suitable location	4.25	0.71	High
2. Instructional facilities	4.50	0.68	High
3. Appropriate duration	4.15	0.81	High
4. Refreshments and prizes	4.20	0.95	High
Average	4.28	0.79	High
Application of Knowledge			
1. Application of knowledge to profession	4.65	0.58	High
2. Confidence and adaptability of knowledge	4.40	0.59	High
3. Ability to share knowledge	4.30	0.92	High
Average	4.45	0.69	High
Total Average	4.18	0.27	High

It could be concluded that the trainees' satisfaction towards the green volunteering activities of students according to the Good-Hearted Vocation Teacher to Support Itinerant Junk Buyers about the workshop sessions on repairing electrical appliances, fixing engines and classifying rubbish was at high level with mean score of 4.18 and S.D. of 0.27.

D. *Results about Self-Assessment of the Workshop Sessions by Students*

TABLE V. SHOWS THE SATISFACTION TOWARDS THE GREEN VOLUNTEERING ACTIVITIES ORGANIZED BY STUDENTS

Item	Mean	S.D.	Level	Rank
Readiness for activities	4.50	0.52	High	5
Step-by-step operation	4.50	0.52	High	5
Task assignment	4.75	0.45	The Highest	3
Collaboration and support	4.83	0.38	The Highest	2
Endeavour	5.00	0	The Highest	1
Information accuracy	4.60	0.51	The Highest	4
Checking errors daily	4.41	0.52	High	6
Mistake correction	4.50	0.38	High	5
Accepting diverse opinions	4.83	0.49	The Highest	2
Creativity in works	4.60	0.49	The Highest	4
Success from activities	4.83	0.38	The Highest	2
Average	4.68	0.42	The Highest	

The results about self-assessment of the green volunteering activities of 13 students were at the highest level with mean score of 4.68 and S.D. of 0.42.

TABLE VI. RESULTS ABOUT THE RECORDINGS AND THE PROBLEMS AS WELL AS SOLUTIONS DURING THE OPERATIONS

Problems found	Solutions discussed
1. Trust from the community members because they thought this project might affect students' grade or their graduation status	Details were given informally to the community members through slideshows and video clips about the SCB Community Challenge Project. The chairperson was elected to act as an intermediary between students and community members. KMUTT has a long experience with research and community and door-knocking approach could help both parties understand and gain trust.
2. The impact of training on their profession each day	The talk was organized between students and community members to find the right time for each training session and the contents would be adjusted to fit the duration.
3. Worries about basic understanding before the training	The training sessions had to be easy and clear with simple exercises before difficult activities. Moreover, slideshows were shown so that the trainees could see and practice the real tools. The public relations were done actively through door-knocking and colorful advertisements.
4. On the first day, children came running around, making workshop sessions hard to manage	Another team was specifically organized for children with the topics about manners and social etiquette, rules and beauty of the world through video clips and games. This could help reduce the worries of many adults who need to take care of their children. More adults attended the training and the teachers could manage their workshop easily without children running around.
5. Delays in each day due to travels and presentations	Devices were prepared in advance before the workshop sessions. Teaching was also prepared before each session through practice with advisors and experienced teachers in each area. Demonstrations were done before hands-on experiences in order to time the whole process. Errors were corrected and good understanding was built among the working group in terms of contents for each day and the age group. For difficult areas, guest speakers would be sought.
6. Location	The sessions on engines needed saleng as the main item for the whole training. The location at outdoor public area was good but the time of training was not suitable because afternoon time was too hot although tents were provided. There might be rain on some days to test the endurance of the speakers and the trainees. However, the working group and the community offered refreshments to cheer them up.
7. Late attendance	Attendance was checked exactly at 1pm and then theories were introduced before the practical section. However, some community members might be back from work later that day but they would like to attend the sessions. The teachers were afraid that they might not catch up and as such a special team was set for one-to-one tutorial to open up opportunities for community members to gain the same understanding.

VIII. SUGGESTIONS

A. *Short-term suggestions*

a) Another team should be set up to consolidate the knowledge of the community members by assigning some people to find faulty products so that they could practice repairing and gain knowledge. This could be helpful for everybody.

b) Additional skills should be provided to increase the expertise and to future advantages.

c) Brokers are needed to buy the goods from the community.

d) Housewives could be given a training session on inventions from unwanted products.

B. *Long-term suggestions*

a) Savings group should be established for disadvantaged groups in Bangkok.

b) Junk banks could be established to trade the products within the community.

IX. ACKNOWLEDGEMENTS

Faculty of Industrial Education and Technology, King Mongkut's University of Technology Thonburi would like to express gratitude to the funding from Siam Commercial Bank PLC.

REFERENCES

[1] KingMongkut's University of Technology Thonburi, 2011, Community Research Project to Reduce and Solve the Social Inequality in Bangkok: A Case Study of Community under the Bridge of Zone 1, Toong-kru District, Bangkok, National Institute of Development Administration. pp. 20 – 32.

[2] Faculty of Industrial Education and Technology , King Mongkut's University of Technology Thonburi , [Online], Available : http://www.fiet.kmutt.ac.th/home/index.php/about faculty,2013, [Retrieved 1 March , 2013].

[3] Energy Environment Safety and Health, King Mongkut's University of Technology Thonburi , [Online], Available : http://www.eesh.kmutt.ac.th/doc/eesh_index_t.asp,2013,[Retrieved5 March ,2013].

[4] SCB Community Challenge 2009 [Online], Available : http://www.scbchallenge.com/challenge2009/community_howto.php , [Retrieved 4 March , 2013].

APPENDIX

Fig. 1. Green Volunteering Activities of Students in a Disadvantaged Area According to the Good-Hearted Vocation Teacher to Support Itinerant Junk Buyers.

Knowledge Management Strategyfor SMEs

Kitimaporn choochote
Faculty of Technology and Environment
Prince of Songkla University Phuket Campus
Phuket, Thailand

Abstract—**In Thailand, as in other developing countries, the focus was on the large industry first, since governments assumed that large enterprises could generate more employment. However, there has been a realization that the SMEs are the biggest group in the country and are significantly important to the process of social and economic development. This realization has prompted Thailand to institute mechanisms to support and protect SMEs consist of manufacturing, merchandising (wholesale & retail) and service businesses. Unfortunately, most of these SMEs lack capability in operational areas such as technology, management, marketing, and finance when compared to large enterprises. In order to adapt and survive SMEs need full and proper support from the government. To aid in their adaptation and survival, SMEs and government must develop their knowledge management framework to effectively harness their past and present experiences, and anticipate the future evolution of their commercial environment. In most countries, SMEs are the biggest source of export even in normal circumstances. Consequently, the state and SMEs have to focus and work hard to their ensure survival.**

Keywords—Knowledge Management; Small and Medium Enterprice; SME;

I. Introduction

While knowledge management is recognized as management of the 21st century, there are many problems if people launch programs of knowledge management without due consideration to factors which facilitate or hinder the knowledge management process. Therefore, understanding the factors affecting success and failure of knowledge management processes is an important key to help managers identify and understand what is required to make knowledge management work. Once the factors are understood, they can develop related context that influences the effectiveness of their knowledge management processes [1].The knowledge diagnostic remains one of the least understood aspects of knowledge management, that is, how central a role knowledge assets, or lack thereof, play in people's capabilities to deliver quality work, or in the enterprise's ability to pursue and achieve strategic goals. Competent knowledge diagnostics rely on an integrated understanding of how competent intellectual work contributed, and how the myriad of knowledge management solution alternatives that are available can help conduct effective and systematic knowledge management [2].

Management scholars and writers, including Nonaka and Takeuchi, Drucker, Leonard-Barton, Senge, Quinn, and more recently, Davenport and Prusak made an impact in the 1990s through influential books with different points of view. Organizations are confused about where and how to start, even if they acknowledge that knowledge management could make a difference to their performance. There is a need for models, frameworks, or methodologies that can both help us to understand the sorts of knowledge management initiatives or investments that are possible and to identify types of knowledge management that make sense in each context.

Even though the utilization of knowledge has become a key factor for the success of organizations, management has found it difficult to transform their firm through programs of knowledge management. Many models and theoretical frameworks from various perspectives try to explain knowledge management, but empirical proof of knowledge related hypotheses are scarce; also there is a lack of coherence between different concepts of knowledge management. For practicing managers, there is a major gap between knowledge management theory and practice. It is therefore essential to gain a clear and comprehensive understanding of how knowledge works within the organization [3]. Research is needed to build a comprehensive model of the context of knowledge management strategy and more so how it applies to SMEs in the developing world. This research will examine knowledge management processes used by SMEs in Thailand and by doing so it will contribute towards the difficult problem of using knowledge management processes within a given context [4].

II. Literature review

A. Knowledge Management View

There are many models which have been developed and published where knowledge management is concerned. However they tend to focus on large enterprises, while this research focuses on small and medium enterprises. Some of the terminology and theories will be borrowed from these models throughout this research. One such concept, developed by Nonaka, is 'ba'. Even though throughout the research this concept may not be specifically referred to, its theme will show up.

B. Knowledge Management Success Stories

Probably one of the most well know documented cases of knowledge management successes can be found in the works of the prominent Japanese author, Ikujiro Nonaka. In his book titled "The knowledge-creating company: how Japanese companies create the dynamics of innovation", Nonaka documents the knowledge management strategies of major Japanese companies in the automobile and electronics industries. Companies such as Honda, Canon, NEC and Nissan are analyzed to investigate the relationship between their knowledge management strategy and their success [5].

Other examples of documented successful cases can be found in the book "Leading with knowledge: Knowledge Management Practices in Global Infotech Companies" by Madanmohan Rao, where prominent companies such as Novell, Oracle and IBM are studied for links between their knowledge management strategies and their successes [6].

C. Small and Medium Enterprice (SMEs) in Thailand

Small and medium enterprises in Thailand are defined according to a regulation passed by the Ministry of Industry in September 2002. The Ministry defines SME by business sector and, within each sector, by the number of employees or value of fixed assets (excluding land). The four business sectors are as follows:

TABLE I. CLASSIFICATION OF SMEs IN THAILAND [7]

	Small		Medium	
Type	No. of Employees	Fixed Assets (THB million)	No. of Employees	Fixed Assets (THB million)
Manufacturing	Not more than 50	Not more than 50	51-200	51 - 200
Services	Not more than 50	Not more than 50	51-200	51 - 200
Wholesale	Not more than 25	Not more than 50	26-50	51 - 100
Retail	Not more than 15	Not more than 30	16-30	31 - 60

Ministry of Industry, 2002

Manufacturing enterprises defined in terms of permanent assets and include the value of the land are classified into:

- Small Enterprises – fixed assets not more than 50 million Bath and number of employees not more than 50
- Medium Enterprises – fixed assets above 50 & up to 200 million Bath and number of employees above 50 & up to 200

Service enterprises defined in terms of permanent assets and include the value of the land are classified into:

- Small Enterprises – fixed assets not more than 50 million Bath and number of employees not more than 50
- Medium Enterprises – fixed assets above 50 & up to 200 million Bath and number of employees above 50 & up to 200

Wholesale enterprises defined in terms of permanent assets and include the value of the land are classified into:

- Small Enterprises – fixed assets not more than 50 million Bath and number of employees not more than 25
- Medium Enterprises – fixed assets above 51 & up to 100 million Bath and number of employees above 25 & up to 50

Retail enterprises defined in terms of permanent assets and include the value of the land are classified into:

- Small Enterprises – fixed assets not more than 30 million Bath and number of employees not more than 15
- Medium Enterprises – fixed assets above 30 & up to 60 million Bath and number of employees above 15 & up to 30

In various countries, either classified as developed or developing ones, the definition and the importance of SMEs are similar. However, the intention in looking for new approaches in order to make SMEs a genuine source of national revenue might be different.

III. RESEARCH METHODOLOGY

This research covers the methodology employed in this study. It includes a description of the sample, sample size and the population, the scope of the study, the data collection methods, tools used and methods of statistical analysis to investigate the research hypothesis presented in chapter one.

A. Population and Sample

The population of the study consisted of Small and medium enterprises (SMEs) which manufacture automobilecomponents in Bangkok, Thailand. The companies which register their company names with the Department of Industrial Works, Ministry of Industrial, Thailand, were considered. The number of registered companies manufacturing automobile components in the entire country is 1,724. According to the Department, the number of companies which manufacture automobile components of SMEs in Bangkok is 430.

The sample was selected based on the willingness of companies to participate in the study. A preliminary meeting was held with the respective Chambers of Commerce in Bangkok to establish willingness and whether companies met the criteria for the study. Based on these meetings, 20 companies were selected from Bangkok.

The sample size of 20 may seem small when compared with a true population size of > 6,000 and 430, however many of the companies which comprise the whole population were not classified as SMEs (either micro or large enterprises) and hence were not suitable for this study. The sample size of 20 from Bangkok was deemed appropriate when this fact was taken into consideration.

B. Construction of the Questionnaire

The questionnaire had seven parts; each part consists of groups of questions as follows:

Part 1Demographic data of the interviewee
Part2 Characteristics of company
Part 3Strategy, management style and IT investment
Part 4 Knowledge management process of the company
Part 5 Customer factors considered by thecompany
Part 6Attitude of the company towards government
Part 7 Private and international organization's support

C. Data Collection

The owners or chief executives of the enterprises were interviewed and the questionnaires were filled during the interview process. The data were collected between the months of April – July, 2008 in Bangkok.

D. Data Analysis

The data were analysed in three stages which are listed

As; Preliminary and summary analysis, Framework analysis, and Correlation analysis.

IV. RESEARCH OBJECTIVGE

Considering the need for research, the objectives for the proposed research are identified as:

To identify a relationship, if any, between theknowledge management processes of these companies and their sales performance.

V. HYPOTHESIS OF THE STUDY

This research intended to test the hypothesis - sales performance of SMEs is related to the knowledge management process adopted by SMEs. Statistically this hypothesis was stated as:

H0 = Sales performance of SMEs is not related to the knowledge management strategy adopted by SMEs.

H1 = Sales performance of SMEs is related to the knowledge management strategy adopted by.

To test this hypothesis two variables were identified. The first variable (V1) was the sales performance of the SMEs in the sample while the second variable (V2) was a measure of the knowledge management strategy used by the SMEs in the sample, based on the parameters investigated by questionnaire which were combined using a model presented in this research.

VI. KNOWLEDGE MANAGEMENT FRAMEWORK

Often SMEs overlook simple solutions which would increase their productivity and competitive advantage. SMEs fail to devote enough attention to technologies and tools such as groupware, data mining, semantic networks, knowledge maps and content management systems which provide the technological foundation for a knowledge management process. Many times it's the case that these tools and technologies may be freely or cheaply available, but because the interest is not there within SMEs these tools are not used [8]. In the book Knowledge Integration it is suggested that SMEs need to work smarter and should spend some effort educating professionals in the use of tools which may potentially tap their knowledge reserves [9]. But in order to use knowledge management software tools the requisite information technology (IT) infrastructure must be in place. Having good IT infrastructure upon which knowledge management software tools can be deployed as well as having company policies which are conducive for knowledge creation and sharing combined with other relevant factors yields a good knowledge management process for SMEs. The diagram in figure 1 summarizes this point and is used as the framework for analyzing the knowledge management strategies of the sample.

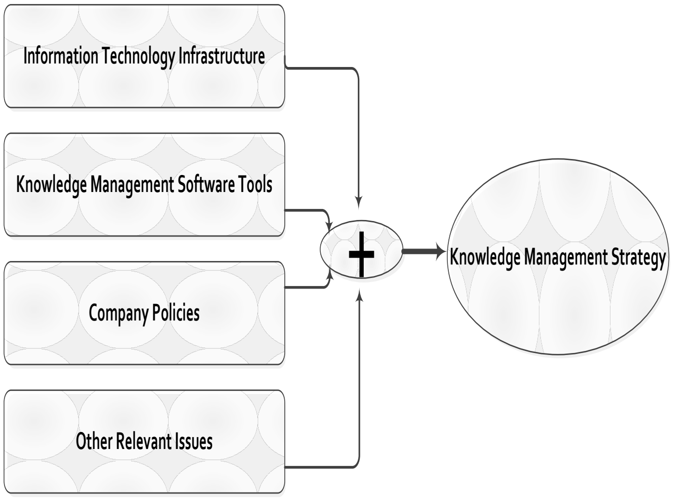

Information Technology Infrastructure
Knowledge Management Software Tools
Company Policies
Other Relevant Issues
+
Knowledge Management Strategy

Fig.1. Knowledge Management Framework for SMEs

The model consists of four components which when effectively combined produces a solid knowledge management strategy for SMEs. Below is an explanation of each component:

A. Information technology infrastructure

This component primarily focuses on the computerhardware, software, storage, and networking setup of the organization. It is the foundation for the knowledge management software tools component of the model. Issues such as operating systems, network speeds and data storage capacity is addressed by this component. It is absolutely essential that the companies' IT infrastructure meets the requirements for running and supporting the knowledge management software tools.

B. Knowledge management software tools

This component comprises of a variety of software and tools; some of which are described in the following list:

- Groupware (collaborative software) - is software designed to help people involved in a common task achieve their goals. Groupware may include software that ranges from simple messaging, emailing and conferencing software to full project management tools [10].
- Data mining software – is software which analyzes data to determine whether useful patterns exist which may be exploited by a company to gain a competitive advantage. There are many open packages, such as the WEKA tool kit or Rapid Information Miner, which may be utilized by a company with little effort or training [11].
- Semantic networks- is a network which represents semantic (refers to meaning) relations between the concepts (and idea) [12].
- Knowledge maps- A knowledge map portrays a perspective of the players, sources, flows, constraints and sinks of knowledge within an organization. It is a

navigation aid to both explicit (codified) information and tacit knowledge, showing the importance and the relationships between knowledge stores and the dynamics. The final 'map' can take multiple forms, from a pictorial display to yellow pages directory, to linked topic or concept map, inventory lists or a matrix of assets against key business processes.[13]

- Content management systems- is concerned with content, documents, details and records related to the organizational processes of acompany. The purpose is to manage the organization's unstructured information content, with all its diversity of format and location [14].

C. Company policies

This component refers to what mechanisms, culture andrules the company has which affects the creation, distribution and management of knowledge.

D. Other relevant issues

This component is a bit dynamic since it can include factors like expert consultation, customer related activities,government support, relevant laws and assistance schemes, private and international organizations support, global market trends and a host of other issues. It is up to the SMEs to be cognizant of their working environment and capitalize on opportunities which may sporadically arise.

VII. Assessment of the Knowledge Management Strategy of the Sample

The sample seemed to demonstrate the superior IT infrastructure. All the companies had at least one IT professional on staff. However, it was found that the sample was neglectful in areas such training and investment in training.

In the sample 90% of the companies used specialized software which supports component two of the analysis framework. However, coupled with the fact that there was little IT training or investment in IT training suggests that only the IT professionals have knowledge of the use of the specialized software and time is not taken to teach other employees. There may be several reasons for this, such as trust issues, however this type of approach is not conducive to a healthy knowledge management strategy.

On the issue of company policies, the sample demonstrated they have the policies in place to support a strong knowledge management process. The sample under utilized private and international organizations services as well as government services. This is a trend that should be remedied once cost is not prohibitive. The companies need to invest time in investigating what services are available which could lead to a competitive advantage.

On the issue of advertising the companies recognized the need for a strong advertising program and this contributes positively to their knowledge management strategy.

All in all the companies have demonstrated a slightly above average knowledge management strategy. However they need to explore the possibility of using government and private and international organizations support as well as look into training other staff in IT technology.

VIII. Analysis of The Knowledge Management strategy of the Sample Against Sales

In this part the model and the results were used to perform correlation analysis between the knowledge management strategy of the SMEs and their previous year's sales performance. This analysis investigated the statistical stated hypothesis:

H0 = Sales performance of SMEs is not related to the knowledge management strategy adopted by SMEs.

H1 = Sales performance of SMEs is related to the knowledge management strategy adopted by.

To analyze whether the knowledge management strategy is indeed a factor which can stimulate an SME's sales performance the following steps were taken:

1) *The attributes which comprised each component of the analysis framework and produced an overall objective component score. Each component score was tested against the previous year's sales performance for correlation. The analysis was performed on the sample as a collective for thoroughness.*

2) *Lastly and most importantly an objective assessment of the knowledge management strategy of each company in the sample were obtained, by combining each of the four components of the framework, and tested for correlation with their sales performance. This analysis was also performed on the sample as a collective for thoroughness.*

Even though correlation does not imply causation [15], using well established works, presented in the literature review of this research, which show good knowledge management strategies create a competitive advantage which translates into profits, it would be reasonable to conclude that at least part of the SME's success is as a result of its knowledge management strategy. If it is shown that the sales performance of the sample was not significantly correlated to its knowledge management strategy, then the question that would be answered by this study is why, in this case, the results differ from theoretical and empirical results from the established literature on knowledge management.

B. Framework Components and Sales Performance Correlation Analysis

TABLE II. Framework Components Correlation with Sales Performance the Sample

No	Parameters	Thailand	
		rho	Sig(2-tailed)
1	Information technology infrastructure	0.403	0.078
2	Knowledge management software tools	-0.304	0.193
3	Company policies	0.378	0.100
4	Other relevant factors	0.124	0.604

** Correlation is significant at the 0.01 level (2-tailed)
* Correlation is significant at the 0.01 level (2-tailed)

Table II.Shows that none of the Framework Components had a 2-tailed level of significance less than or equal to 0.05 when tested for correlation with sales performance. This observation indicated that none of the components were significantly correlated with the sales performance of the sample.

C. Knowledge Management strategy Estimate and Sales Performance Correlation Analysis

TABLE III. KNOWLEDGE MANAGEMENT STRATEGY ESTIMATE CORRELATION WITH SALES PERFORMANCE THE SAMPLE

No	Parameters	Thailand	
		rho	Sig(2-tailed)
1	Knowledge management process estimate	0.411	0.072

** Correlation is significant at the 0.01 level (2-tailed)
* Correlation is significant at the 0.01 level (2-tailed)

Table III.Shows that the knowledge management strategiesestimate did not have a 2-tailed level of significance less than or equal to 0.05 when tested for correlation with sales performance. This observation indicated that the knowledge management strategy estimate was not significantly correlated with the sales performance of the sample.

IX. CONCLUSIONS, EXPLANATIONS AND IDEAS

This provides explanations, conclusions and ideas for the significant results obtained and the significant results following:

1) The individual attributes which had a 2-tailed level of significance less than or equal to 0.05 when correlated with the sales performance of the respective samples.

2) The framework components correlation analysis with the sales performance of the respective samples.

3) Most importantly, the knowledge management strategy estimate correlation analysis with the sales performance of the respective samples. Validation of the hypothesis occurs in this section.

B. Explanation of the significantly correlated individual attributes

TABLE IV. SIGNIFICANT CORRELATION WITH SALES PERFORMANCE FROM THE SAMPLE

Parameters	Spearman rank correlation coefficient	Sig.(2-tailed)
Investment in IT Infrastructure.	0.628	0.003
Your knowledge management process gives advantages to your company.	0.487	0.030
An application of your knowledge management process helps your management	0.487	0.030

** Correlation is significant at the 0.01 level (2-tailed)
* Correlation is significant at the 0.01 level (2-tailed)

Table IV. Shows that there were three attributes which demonstrated noteworthy 2-tailed levels of significance. Based on the 2-tailed significance of Spearman rank correlation coefficient (rho) between the investment in IT infrastructure and sales performance for Thai companies it has been established that there is a positive relationship between the two variables. This same relationship was dealt with for the combined sample and the inferences and thoughts expressed while analyzing that phenomenon are equally applicable here.

The responses to the statements "your knowledge management strategy gives advantages to your company" and "an application of your knowledge management strategy helps your management" were identical and will be analyzed together. Based on the 2-tailed level of significance of the Spearman rank correlation coefficient between these statements and sales performance it has been established that there is a positive relationship between each statement and the sales performance. From this observation the inference can be made that these two aspects of the Thai sample's perception of their knowledge management strategy were reflected in their sales performance. However, all other perceptions were not reflected in their sales performance. This phenomenon indicates that there may be some facets of their perception of the knowledge management strategy which contributes positively to their sales.

C. Framework components analysis with sales performance

TABLE V. FRAMEWORK COMPONENTS CORRELATION WITH SALES PERFORMANCE FROM THE SAMPLE

Parameters	Spearman rank correlation coefficient	Sig.(2-tailed)
Information technology infrastructure	0.403	0.078
Knowledge management software tools	-0.304	0.193
Company policies	0.378	0.100
Other relevant factors	0.124	0.604

** Correlation is significant at the 0.01 level (2-tailed)
* Correlation is significant at the 0.01 level (2-tailed)

Table V. Shows that none of the Framework Components had a 2-tailed level of significance less than or equal to 0.05 when tested for correlation with sales performance. Based on this observation it can be concluded that none of the framework components were correlated to sales performance of the sample.

D. Knowledge management process estimate analysis with sales performance

TABLE VI. KNOWLEDGE MANAGEMENT PROCESS ESTIMATE CORRELATION WITH SALES PERFORMANCE

Parameters	Thailand	
	rho	Sig(2-tailed)
Knowledge management strategy estimate correlation with sale performance	0.411	0.072

** Correlation is significant at the 0.01 level (2-tailed)
* Correlation is significant at the 0.01 level (2-tailed)

As seen from table VI. Above, the observed 2 tailed level of significance when the knowledge management strategy estimate was analyzed with the sales performance of the sample, using the Spearman rank correlation coefficient (rho) was 0.072. This observation provides statistical evidence to accept H0. Which meant variable 1 (V1 – sales performance) and variable 2 (V2 – knowledge management strategy estimate) were not related. Hence H0 which states sales performance of SMEs is not related to the knowledge management strategy adopted by SMEs in Thailand was validated. Therefore the null hypothesis has been statistically tested and validated.

E. Insights for results

This research has in this instance statistically validated the hypothesis – H0: sales performance of SMEs is not related to the knowledge management strategy adopted by SMEs in developing countries. There are two factors among others which standout as an explanation for the results.

The first and most influential factor is the SMEs' understanding of the knowledge management strategy. Knowledge management and its potential benefits are still in its infancy stages in developing countries. Due to this immaturity there is significant naivety in the understanding and implementation of knowledge management strategies in the context of SMEs. The samples' perception of their knowledge management strategies was adequate, but their perceptions were not reflected in their sales performance. This mismatch is attributed to the fact that their understanding and thus the implementation of the knowledge management strategy was flawed. From this conclusion the recommendations of this research are abundantly applicable.

The second factor which most likely had an effect on the results of the study was the state of the economy at the time of conducting the study. There was a boom in the auto components industry. The economic climate created a condition where manufacturers could sell their products and services without having to invest in the machinations which produce a competitive advantage. This type of climate obscured the weaknesses in the SMEs' management and knowledge management strategy and thus created the perception that the knowledge management strategy was functioning effectively.

REFERENCES

[1] Gold, Andrew H., Malhotra, Arvind and Segars, Albert H. "Knowledge Management: An Organizational Capabilities Perspective," Journal of Management Information Systems, 18.1 (2001), pp. 185-214.

[2] Wiig, Karl M, "The knowledge," Inside Knowledge 6.2 (25 September. 2002),available online at www<http://www.ikmagazine.com/xq/asp/ sid. 0/articled.FD0C4391-F53F-4F27-8B45-CE89C7EEC7D/e Title.The_knowledge_Karl_Wiig/qx/display.htm>.

[3] Reinhardt, Rudiger, "Knowledge Management: Linking Theory to Practice. In: Knowledge Management: Classic and Contemporary Works," Daryl Morey, Mark Maybury and Bhavani Thuraisingham, India: Universities Press, 2001, pp. 187-221.

[4] K Choochote. " An Analysis of Knowledge Management Process for SMEs in Developing Countries: A Case Study of SMEs in India and Thailand." International Journal of Information and Education Technology (2012): 239-242. Print.

[5] Nonaka, Ikujiro, "The Knowledge-Creating Company (Harvard Business Review Classics)," Boston: Harvard Business School Press, December 2008, pp. 23-56.

[6] Rao, Madanmohan, "Leading with knowledge: Knowledge Management Practices in Global Infotech Companies," New Delhi: Tata McGraw-Hill Publishing Company, 2003, pp. 339-532.

[7] Ministry of Industry Thailand, Office of Small and Medium Enterprise Promotion. Definition of SME in Thailand. OSMEP., 2002. available onlineat www.http://cms.sme.go.th/cms/c/portal/layout?p_l_id=47.105

[8] K Choochote and R Nurse. "A Simple Knowledge Management Strategy Model for SMEs in Developing Countries." World Academy of Science, Engineering and Technology. International Journal of Medical and Biological Engineering, April 11-13, Venice: Italy, 2012. 189-192. Print.

[9] Wijnhoven, "Knowledge integration [electronic resource]: the practice of knowledge management in small and medium enterprises," Springer, 2006, pp. 2.

[10] Rao, Madanmohan, " Knowledge Management Tools and Techniques: Practitioners and Experts Evaluate KM Solutions," New Delhi: Elsevier Inc., 2006, pp. 9-12.

[11] Witten, Ian H., and Frank, Eibe, " Data Mining: Practical Machine earning Tools and Techniques, Second Edition," San Francisco: Morgan Kaufmann, 2005, pp. 4-9.

[12] Deliyanni, Amaryllis., and Kowalski, Robert A, "Logic and Semantic Networks: Communications of the ACM, Artificial Intelligence/ Language Processing," March, 1979, pp. 184-192.

[13] Subrt, T., and Brozova, H, "Knowledge Maps and Mathematical Modelling." The Electronic Journal of Knowledge Management," Vol.5 Issue 4, pp. 497 - 504, available online at www.ejkm.com.

[14] Bluebill, ADV, "The Classification & Evaluation of Content Management Systems," The Gilbane Report 11.2 (March, 2003), pp. 2-5.

[15] Myers, Jerome L., and Well, Arnold D, " Research Design and Statistical Analysis," 2nd ed. Mahwah, New Jersey: Lawrence Erlbaum, 2003, p. 508.

A Proposed NFC Payment Application

Pardis Pourghomi
School of Information System,
Computing and Mathematics
Brunel University
London, UK

Muhammad Qasim Saeed
Information Security Group (ISG)
Royal Holloway University of
London Egham, UK

Gheorghita Ghinea
School of Information System,
Computing and Mathematics
Brunel University
London, UK

***Abstract*—Near Field Communication (NFC) technology is based on a short range radio communication channel which enables users to exchange data between devices. With NFC technology, mobile services establish a contactless transaction system to make the payment methods easier for people. Although NFC mobile services have great potential for growth, they have raised several issues which have concerned the researches and prevented the adoption of this technology within societies. Reorganizing and describing what is required for the success of this technology have motivated us to extend the current NFC ecosystem models to accelerate the development of this business area. In this paper, we introduce a new NFC payment application, which is based on our previous "NFC Cloud Wallet" model [1] to demonstrate a reliable structure of NFC ecosystem. We also describe the step by step execution of the proposed protocol in order to carefully analyse the payment application and our main focus will be on the Mobile Network Operator (MNO) as the main player within the ecosystem.**

Keywords—Near Field Communication; Security; Mobiletransaction; GSM authentication.

I. INTRODUCTION

During the past decade, the concept of contactless card technology was introduced to be used in transport, ticketing and in retail. The technology helps people save time by just holding their contactless cards against a reader in a close proximity instead of having to insert the paper cards in and taking it out of the train entrance gates for example. With NFC technology, mobile phones can have additional functionality to act as a contactless card to be used as an easy method of payment. Successful development of NFC technology has recently started in some countries where companies offer several services based on the contactless card technology and mobile phones. Although this technology is increasingly becoming mainstream, it still has issues that need to be addressed [2]. These issues are mainly security concerns with Secure Element (SE) personalization, management, ownership and architecture that can be exploitable by attackers to delay the adaption of NFC within societies. The purpose of this paper is to extend Pourghomi's[1] - this model will be referred as "NFC Cloud Wallet" in this and our future papers - by proposing a complete transaction mechanism based on NFC and GSM networks.

This model is based on cloud computing for the management of payment applications in secure element within the NFC ecosystem. The details of this model are described in Section IV of this paper. As the authentication mechanism of our extended model is based on GSM, we will discuss the GSM authentication later in this paper. We also aim to accelerate the development of NFC mobile payment services by describing the NFC ecosystem in order to raise the attention of business players in terms of the new potential models that can be implemented in order to achieve a cost beneficial and less complex ecosystem framework.

Our contribution in this paper is to extend the NFC Cloud Wallet model and to provide a complete transaction solution based on this model. We propose a model based on the assumption that the cloud is being managed by the MNO.We used the existing security features of GSM network to achieve authentication, data integrity and data confidentiality.

In our proposed model, the SIM is the secure element which is being managed by the MNO. By using our model, a customer with a NFC enabled mobile phone can pay through his cell phone in a secure way.

This paper is organized as follows. Section II consists of a brief introduction to NFC ecosystem with its main elements and functionalities. Section III describes the roles of Secure Element (SE) and the Universal Integrated Circuit Card (UICC) within the NFC ecosystem. Section IV evaluates the previously proposed NFC Cloud Wallet model. Section V discusses GSM authentication that is used in our extended model. Section VI introduces the proposed transaction model as well as the proposed transaction protocoland describes its execution process in details. Section VII is the analysis of our proposed model from multiple security aspects. This analysis encompasses the authentication and security of the messages among customer, shop POS terminal and the MNO. Finally, Section VIII presents our conclusion.

II. NFC & NFC ECOSYSTEM

This section describes the functions of adding the contactless card to mobile phones which produce an intelligent device that enables us to make payments with. This intelligent device is called a "NFC Mobile Phone". When different functions of a mobile phone combine with the functions of contact-less cards, the results of this combination will have a greater significance than just the importance of adding two devices together. This significance defines the NFC-enabled mobile phone which can connect with another NFC-enabled device (i.e. PDA, tablets, etc.) in a short range communication channel. NFC technology enables users to benefit from new and countless services on a daily basis where they can pay for their food; buy a cinema ticket by scanning their phone against a movie poster and much more. This newly developed intelligent device is proposed as an all-in-one personal device that can be personalized and used in a highly interactive environment [3]. Fig. 1 demonstrates the concept of the NFC mobile phone [4].

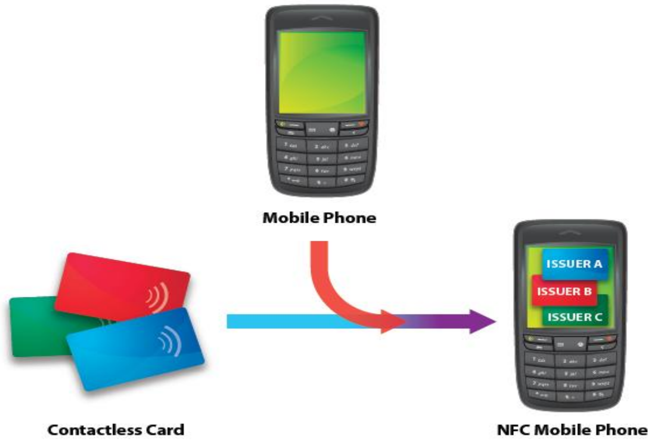


Fig. 1. The concept of NFC mobile phone

The success of the NFC mobile ecosystem is based on the relationships between the involved parties where those relationships have to be clearly defined. The present contactless ecosystem models functionalities can be extended by a well-defined NFC ecosystem which improves the number of functionalities that an involved party can provide. Table I describes the key functionalities of NFC ecosystem [4].

TABLE I. KEY FUNCTIONALITIES OF NFC ECOSYSTEM

Key functionalities	**Description**
Service provisioning	It provides authentication and remote user management due to network availability. Also users can subscribe and personalize their contactless cards.
Mobile network provisioning	It offers user authentication and user care for data connectivity as well as ensuring that the network infrastructure is maintained to enable users to receive data connectivity service.
Trusted Service Manager (TSM)	Delivers a communication platform between Service Providers (SPs) and NFC mobile phones where SPs provide multi-application management functionalities to NFC enabled mobile phone through this platform.

III. SECURE ELEMENT (SE)

The security of NFC is supposed to be provided by a component called security controller that is in the form of a SE. The SE is an attack resistant microcontroller more or less like a chip that can be found in a smart card [3].

SE provides storage within the mobile phone and it contains hardware, software, protocols and interfaces. SE provides a secure area for the protection of the payment assets (e.g. keys, payment application code, and payment data) and the execution of other applications. In addition, SE can be used to store other applications which require security mechanisms and it can also be involved in authentication processes. To be able to handle all these, the installed operating system has to have the capability of personalizing and managing multiple applications that are provided by multiple SPs preferably Over-The-Air (OTA). Still the ownership and control of SE within NFC ecosystem may result in a commercial and strategic advantage but some solutions are already in place and researchers are developing new models to overcome this problem. Universal Integrated Circuit Card is (UICC) is one of the most reliable components to act as a SE in NFC architecture [5]. It is removable, provides the same security as a smartcard, can run multiple applications issued by multiple providers, it is compliant with all smart card standards and it supports GSM and UMTS network. According to GSMA guidelines, UICC is the most appropriate NFC Secure Element in mobile phones [3].

A. *SE Lifecycle*

The **Initialization** of an SE can be completed by different SE issuers such as credit card companies, Mobile Network Operator (MNO), financial institution or retailers. The SEI can also act as a platform provider. If the SE does not contain any applications when issued that means there is no platform manager assigned to that SE. A platform manager cannot deal with SE applications without having different certifications (i.e. Visa PayWave certification).

The *Activation* process takes place when the SE is inserted into the phone. The SE then sings in to the NFC controller and NFC controller sends a confirmation message to the platform manager in order to inform the platform manager of the successful insertion of the SE in the phone. The platform manager then sends a confirmation message to the mobile phone in order to activate the SE. The platform manager is the only party that has the authority to hold the SE keys for data configuration purposes. NFC controller’s identifier is also stored in the SE to inform SE in case if it was inserted into another phone.

During phase 1 of the *Applications Upload* process, the SP (in this case also the Application Issuer (AI)) contacts the MNO that is the only party who is in charge of the Mobile Station International ISDN Number (MSISDN). The only way to classify the external party for an Over-The-Air (OTA) transaction with the NFC phone is the MSISDN.

In phase 2, MNO forwards the SP request to the platform manager (s) that is in charge of the SE. If the there is no SE in the phone, the MNO will inform the SP regarding this issue.

In this case application upload process terminates. But if the platform manager is positive with the request, it will send an offer directly to the SP to upload its application.

In the next phase, SP selects one platform manager amongst others (if more than one platform manager exists) to load its data to the security domain area which is under the control of the same platform manager.

The *Deactivation*procedures are also managed by the platform manager where it can deactivate the SE, OTA in the case of theft or loss. If SE is installed in a new device, then the activation process should be renewed and the platform manager is the only party that should confirm the activation process to enable the SE to be used for contactless transactions [5].

IV. NFC Cloud Wallet Model

This model brought the idea of using cloud computing in order to manage the NFC payment applications which resulted in flexible and secure management, personalization and ownership of the applications [1].This architecture provides easy management of multiple users and delivers personalized contents to each user. It supports intelligent profiling functions by managing customized information relevant to each user in certain environments which updates the service offers and user profiles dynamically. Depending on the MNO network's reception, deployment of this service takes around one minute and deployments can be scaled to any number of users.

The idea of this approach is that every time the customer makes a purchase the payment application which contains the customer's credentials is downloaded into the mobile device (SE) from the cloud and, after the transaction, it is deleted from the device and the cloud will update itself to keep a correct record of customer's account balance. Fig. 2 illustrates the steps that should be undertaken to complete the transaction process [1].

The execution of the model is described in what follows:

1) Customer waves the NFC enabled phone on the POS terminal to make the payment

2) The payment application is downloaded into customer's mobile phone SE.

3) The reader communicates with the cloud provider to check whether the customer has enough credit or not.

4) Cloud provider transfers the required information to the reader.

5) Based on the information which was transferred to the reader, the reader either authorizes the transaction or rejects customer's request.

6) Reader communicates with the cloud to update customer's balance **-** *if customer's request was authorised, the amount of purchase will be withdrawn from his account otherwise customer's account will remain with the same balance.*

As an addition to this model, we suggest that when the NFC enabled phone sends a request to its cloud provider to get permission to make a payment (step 1), the cloud provider sends a SMS requesting a PIN number to identify the user of the phone **-** this is how cloud provider ensures the legitimacy of the phone user. For verification purposes, the customer sends the PIN back to the cloud provider as an SMS.

In order to extend this model, there are two possible approaches to follow. Firstly, the financial institution can be the cloud owner from which the payment application can be downloaded from/into the customer's mobile device; MNO can be linked to the financial institution (that is the cloud owner in this case) or it can stand as a separate party. Secondly, the financial institution could have a contract with a third party company such as PayPal that has its own cloud infrastructure (MNO can be linked with them, it also can stand as a separate party) or the financial institution uses other company's cloud service such as IBM, Microsoft, etc (MNO can be linked with either financial institution, cloud provider or it can stand as a separate party).

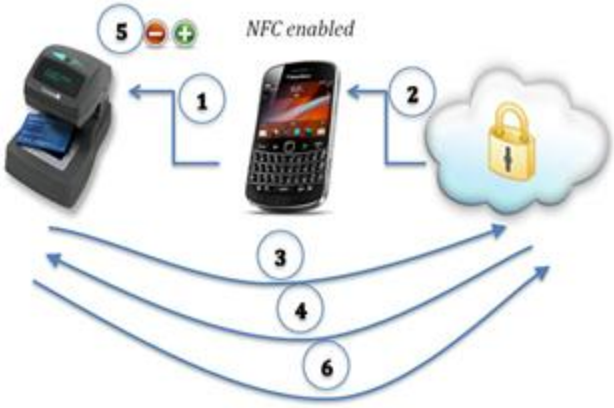


Fig. 2. NFC cloud wallet

V. GSM Authentication

When a mobile device signs into a network, the Mobile Network Operator (MNO) first authenticates the device (specifically the SIM). The authentication stage verifies the identity and validity of the SIM and ensures that the subscriber has authorized access to the network. The Authentication Centre (AuC) of the MNO is responsible for authenticating each SIM that attempts to connect to the GSM core network through Mobile Switching Centre (MSC). The AuC stores two encryption algorithms A3 and A8, as well as a list of all subscribers identity along with corresponding secret key K_i.

This key is also stored in the SIM. The AuC first generates a random number known as *R*. This R is used to generate two responses, signed response *S* and key K_c as shown in Fig. 3, where $S = E_{Ki} (R)$ using A3 algorithm and $K_c = E_{Ki} (R)$ using A8 algorithm [6][7][8][9].

The triplet (R, S, K_c) is known as Authentication triplet generated by AuC. AuC sends this triplet to MSC. On receiving a triplet from AuC, MSC sends *R* (first part of the triplet) to the mobile device. SIM of the mobile device computes the response *S* from *R*, as K_i is already stored in the SIM. Mobile device transmits S to MSC. If this *S* matches the *S* in the triplet (which it should in case of a valid SIM), then the mobile is authenticated.K_c is used for communication encryption between the mobile station and the MNO. Table IIdescribes the abbreviations used in the proposed protocol.

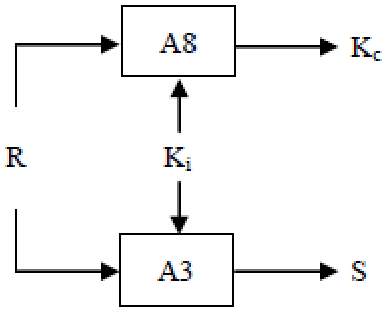


Fig. 3. Generation of *Kc* and *S* from *R*

TABLE II. ABBREVIATIONS

AuC	Authentication Centre (subsystem of MNO)
IMSI	Internet Mobile Subscriber Identity
K_i	SIM specific key. Stored at a secure location in SIM and at AuC
K_c	E_{ki} *(R)* using A8 algorithm
K_{c1}	$H(K_c)$. Used for MAC calculation
K_{c2}	$H(K_c.)$. Encryption key
K_p	Shared key between PG and shop POS terminal
LAI	Local Area Identifier
MNO	Mobile Network Operator
NFC	Near Field Communication
PI	Payment Information
POS	Point Of Sale. Part of shop
R	*RAND*. Random Number (128 bits)
R_s	Random number generated by SIM (128 bits)
TC	Transaction Counter
TRM	Transaction Request Message
TI	Transaction Information
TMSI	Temporary Mobile Subscriber Identity
TP	Total Price
TS_U	User's Time Stamp
TS_{Tr}	Transaction Time Stamp
SD	Shopping Details

VI. Proposed Model

We propose an extension to previously proposed NFC Cloud Wallet model. Since there are multiple options applicable to this model, we designed our model based on the following assumptions:

- SE is part of SIM
- Cloud is part of MNO
- MNO is managing SE/SIM
- Banks, etc. are linked with MNO

These assumptions are appropriate regarding the NFC execution process and its ecosystem. As mentioned in Section IIIpreviously, SE is in the format of UICC therefore SE is part of the SIM. MNO manages the cloud infrastructure and it is the only party that has full access and permission to manage confidential data which are stored in the cloud. As MNO is the owner of the cloud, it fully manages the SIM in terms of monitoring the GSM network and controlling cloud's data. From the financial institution's point of view, they only deal with MNO as MNO is the single party that has full control over the SIM as well as the cloud.

A. The Proposed Protocol

Our proposal is based on cloud architecture where the cloud is being managed by the MNO. The cloud and the banking sector are the subsystems of MNO in our proposal, in addition to the existing subsystems of an MNO. We assume that the communication is secure between various subsystems of the MNO. The shop POS terminal, registered with one or more MNO, shares an MNO specific secret key K_p with the corresponding MNO. This key is issued once a shop is registered with the MNO. The bank detail of the shopkeeper is also registered with the MNO for monetary transactions. The communication between the shop POS terminal and the mobile device is wireless using NFC technology. The mobile device has a valid SIM. We used the existing feature of GSM network for mutual authentication. A recent study by reference [10] proposed a mechanism for GSM authentication in NFC environment. We tailored their model according to our requirement in our proposed architecture. The detailed execution of our protocol is described in Fig. 4.

The proposed protocol executes in three different phases: Authentication, Keys generation and Transaction. The protocol initiates when the customer places his cell phone for the payment after agreeing to the total price displayed on the shop POS terminal. The details of these phases are described in what follows:

B. Phase 1. Authentication

Step 1: As soon as the user places his mobile device, NFC link between the mobile device and the shop POS terminal is established. The shop POS terminal sends an *ID* Request message to the mobile device.

Step 2-3: The mobile device sends *TMSI, LA*I as its*ID*. On receipt of the information from the mobile device, the shop POS terminal determines the user's mobile network. The network code is available in *LAI* in the form of Mobile Country Code (*MCC*) and Mobile Network Code (*MNC*). An *MNC* is used in combination with *MCC* (also known as a *'MCC/MNC tuple'*) to uniquely identify a mobile phone operator/carrier [11].

Step 4-5: The shop POS terminal sends *TMSI*, *LAI*, and Shop *ID* to respective MNO for customer authentication and shop identification.

Step 5.1: In case of incorrect *TMSI*, a declined message is sent.

Step 6: In case of correct identification, the MNO generates one set of authentication triplet (R, S, K_c) and sends R to mobile device through shop POS terminal.

Step 7-8: SIM computes K_c from R as explained in Section V. SIM generates a random number R_s and concatenates with R, encrypts with key K_c and sends it to the MNO through shop POS terminal.

Step 9-10: The MNO checks the validity of the SIM (or mobile device). It receives $E_{Kc}(R||R_s)$ from the mobile device and decrypts the message by K_c, the key it already has in authentication triplet. The MNO compares R in the authentication triplet with the R in the response. In case they do not match, a ‘Stop’ message is sent to the mobile device and the protocol execution is stopped. If both R are same, then the mobile is authenticated for a valid SIM. In this case, the MNO swaps R and R_s, encrypts with K_c and sends it to mobile device.

Step 11-12: This step authenticates the MNO to the mobile device. The mobile device receives the response E_{Kc} ($R_s||R$) and decrypts it with the key K_c already computed in Step 7. The mobile device compares both R and R_s. If both are same, then the MNO is authenticated and a *‘successful authentication’* message is sent to the MNO.

C. Phase 2. Key Generation and PIN Verification

Step 13-14:K_p is a shared secret between the MNO and the shop POS terminal. K_c is the shared secret between the MNO and the customer's mobile device (computed in step 7). There is no shared secret between the POS terminal and the mobile device till this stage. MNO and mobile device compute one-way hash function of K_c to generate K_{c1}, the key that will be used for MAC calculation. The MNO computes K_{c2} from K_{c1} using one-way hash function and sends it to shop POS terminal by encrypting it with K_p. Mobile device also computes K_{c2} as it already has K_{c1} . K_{c2} is the encryption key between MNO, shop POS terminal and the customer's mobile device.

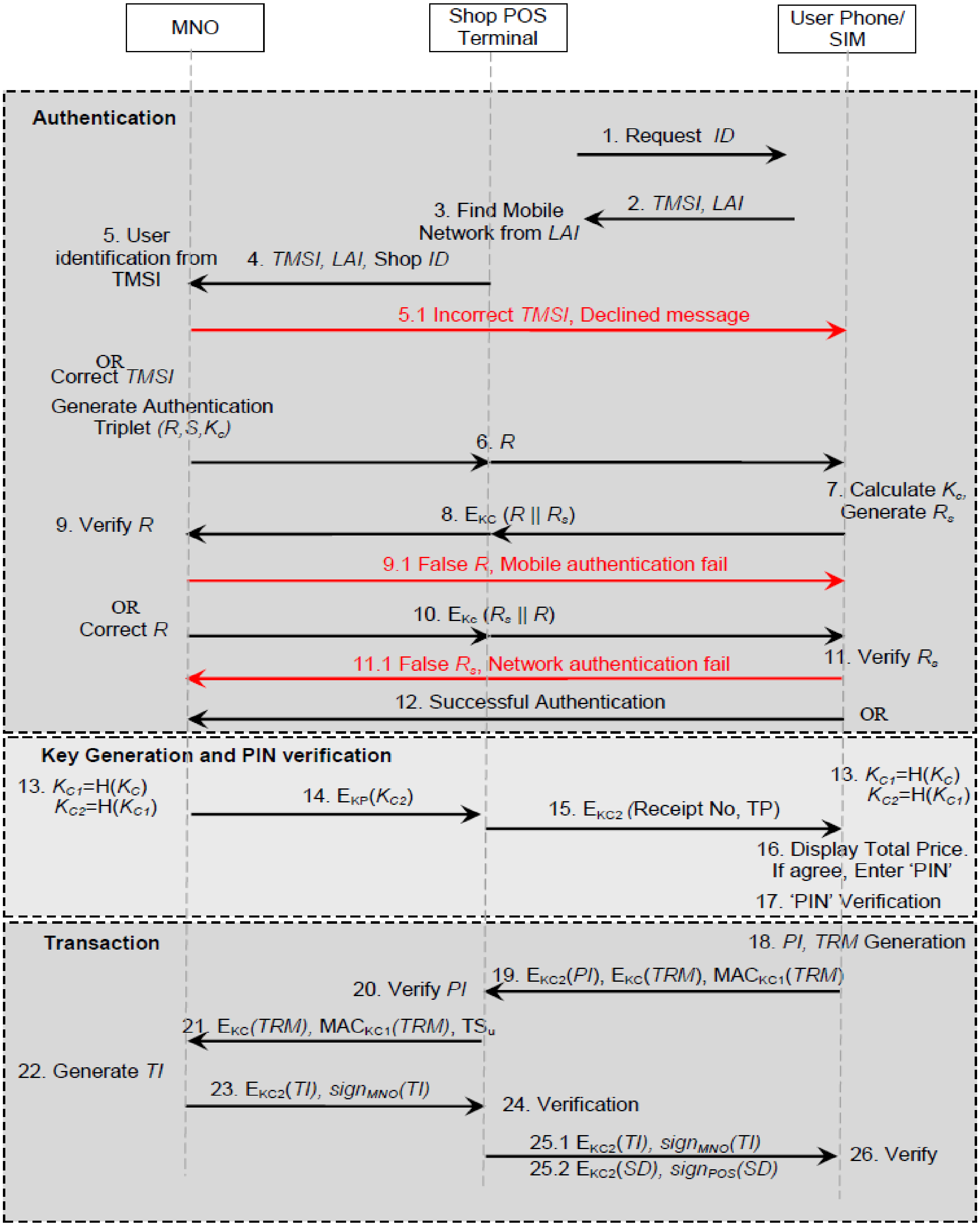

MNO
Shop POS Terminal
User Phone/ SIM
Authentication
1. Request *ID*
2. *TMSI, LAI*
3. Find Mobile Network from *LAI*
4. *TMSI, LAI*, Shop *ID*
5. User identification from TMSI
5.1 Incorrect *TMSI*, Declined message
OR
Correct *TMSI*
Generate Authentication Triplet *(R,S,K_c)*
6. *R*
7. Calculate K_c, Generate R_s
8. E_{KC} (R || R_s)
9. Verify *R*
9.1 False *R*, Mobile authentication fail
OR
Correct *R*
10. E_{Kc} (R_s || R)
11. Verify R_s
11.1 False R_s, Network authentication fail
12. Successful Authentication
OR
Key Generation and PIN verification
13. K_{C1}=H(K_C) K_{C2}=H(K_{C1})
14. E_{KP}(K_{C2})
15. E_{KC2} (Receipt No, TP)
13. K_{C1}=H(K_C) K_{C2}=H(K_{C1})
16. Display Total Price. If agree, Enter 'PIN'
17. 'PIN' Verification
Transaction
18. *PI*, *TRM* Generation
19. E_{KC2}(*PI*), E_{KC}(*TRM*), MAC_{KC1}(*TRM*)
20. Verify *PI*
21. E_{KC}(*TRM*), MAC_{KC1}(*TRM*), TS_u
22. Generate *TI*
23. E_{KC2}(*TI*), $sign_{MNO}$(*TI*)
24. Verification
25.1 E_{KC2}(*TI*), $sign_{MNO}$(*TI*)
25.2 E_{KC2}(*SD*), $sign_{POS}$(*SD*)
26. Verify

Fig. 4. The proposed protocol

Step 15-17: The shop POS terminal sends the Total Price (*TP*) and the Receipt Number encrypted with K_{c2}. The user's mobile device decrypts the information and displays to the user. If he agrees, he enters the PIN. The PIN is an additional layer of security and adds trust between the user and the shopkeeper. A PIN binds a user with his mobile device, so the shopkeeper is to believe that the user is the legitimate owner of the mobile device. Moreover, the user feels more secure as no one else can use his mobile device for transaction without his consent. PIN is stored in a secure location in the SIM. The SIM compares both PINs and if both are same, the user is authenticated as the legitimate user of the mobile device. Otherwise, the protocol is stopped.

D. Phase 3. Transaction

Step 18: The customer's cell phone generates two messages, *PI* and *TRM*, such that;

$$PI= Receipt\ No,\ Total\ Price,\ Time\ Stamp\ (TS_U)$$
$$TRM=PI,\ R_s,\ Transaction\ Counter$$

Step 19:TS_U represents the exact time and date the transaction has been committed by the user. *TC* is a counter that is incremented after each transaction and is used to prevent replay attack. *PI* is encrypted with K_{c2} so that it can be verified by the shop POS terminal. The user encrypts the *TRM* with K_c so that it cannot be modified by the shop terminal. The user computes MAC with K_{c1} over the *TRM* using Encrypt-then-MAC approach for integrity protection.

Step 20-21: The POS terminal can decrypt only the *PI* encrypted with by K_{c2} to check its correctness. The POS terminal does not need to verify the MAC (and it cannot do so), as it already knows the main contents of *PI*. The Shop POS terminal also verifies the TS_U to be in a defined time window. If *PI* is correct, the POS terminal relays the encrypted *TRM* with corresponding MAC along with the TS_U to the MNO.

Step 22: On receipt of the message, the MNO checks the integrity of the message by verifying the MAC with K_{c1}. If the MAC is invalid, the transaction execution is stopped. In case of a valid MAC, the MNO decrypts the message. The MNO compares the R_s in the *TRM* with the *Rs* received earlier in the authentication phase. A correct match confirms that the user is the same who was earlier authenticated.It also verifies the *TC* and TS_U. In case of successful verification, the MNO communicates with the concerned subsections for monetary transaction. The concerned subsections of the MNO checks the credit limitations of the user, and if satisfied, executes the transaction. Once the transaction is executed, the MNO generates *Transaction Information (TI)* message as:

$TI = Transaction\ Serial\ Number,\ Amount,\ TS_{Tr}$

Step 23-25: The MNO encrypts *TI* with K_{c2}, digitally signs the message and sends it to the shop POS terminal. The POS terminal verifies the signature. A valid signature indicates correct *TI*. The POS also verifies the *TI* for the amount mentioned in the *TI*. In case of successful verification, the POS terminal appends the message it received from the MNO with the *Shopping Details (SD)* and corresponding digital signature.

Step 26: The user verifies both signatures. It verifies the contents of *TI* and *SD*.

VII. Protocol Analysis

In this section, we analyse our proposed model from multiple security aspects. This analysis encompasses the authentication and security of the messages among customer, shop POS terminal and the MNO. The analysis also includes multiple attack scenarios, such as a customer is dishonest and has intentions to pay less, or the shopkeeper is dishonest and has plans to receive more money.

A. Mutual Authentication

A mutual authentication between a customer and MNO occurs whenever the customer agrees to pay some amount. Since this authentication is performed through shop POS terminal, we analysed our protocol from an angle that if the POS terminal has some malicious intentions. In this case, there can be following two scenarios:

1) POS Terminal Impersonation as a Customer

We assume that the shop POS terminal is dishonest and keeps a record of all messages against a legitimate customer (we call it as 'target customer'). The aim of the shopkeeper is to transfer money from the target customer without his consent. The shop POS terminal impersonates as target customer to the MNO by replaying message 4. In case the *TMSI* and *LAI* are valid at that time (the chances are higher if the message is replayed just after the legitimate transaction of the target customer), the MNO will send a random number *R* to the terminal. *R* is 128 bit random number generated by the MNO so the chances for its repetition are almost negligible. The shopkeeper cannot compute a valid response in step 8 for a different *R*, as the shop lacks K_i to compute K_c. Therefore, a shop cannot successfully impersonate as a customer by replaying old messages.

2) POS Terminal Impersonation as MNO

In this scenario, we assume that the shop is dishonest and communicates with a target customer without establishing a communication link with the MNO. Again, we assume that the shop keeps a record of legitimate messages of the target customer. The shop sends message 1 (*Request ID*) to the target customer and gets its response in message 2. Since shop does not communicate with MNO in this scenario, it does not send message 4 to MNO. However, the shop replays the recorded *R* in message 6 to the target customer. The target customer believes that he has been correctly identified by the MNO and the *R* is actually generated the MNO. So the user computes a response and sends it in message 8 to the shop. Message 8 contains R_s encrypted with the K_c. The R_s is a random number generated by the SIM and is different in each transaction. So, message 8 will be different than the one already recorded with the shop. Since message 8 is different, the shop can neither replay message 10, as it will be different for this transaction, nor it can compute a valid message 10. This scenario is, again, not successful.

B. Encryption and MAC Keys

Separate keys are used for encryption and MAC calculation making the protocol more secure. *Encrypt-then-MAC* is an approach where the ciphertext is generated by encrypting the plaintext and then appending a MAC of the encrypted plaintext. This approach is cryptographically more secure than other approaches [12]. Apart from cryptographic advantage, the MAC can be verified without performing decryption. So, if the MAC is invalid for a message, the message is discarded without decryption. This results in computational efficiency.

C. User Interaction

The user interaction with the system is reduced to single interaction making it a user-friendly protocol. The user feels more secure as the transaction is protected by PIN verification. There are chances that a user withdraws his mobile device from NFC terminal as a psychological move to enter PIN. This will break NFC link, but as the PIN is stored in the SIM, it does not require NFC link for verification. Once the user PIN has been verified by the SIM, the user places his mobile device back on the NFC terminal and the protocol resume from the same point. There are chances that a dishonest user withdraws his mobile device in order to enter the PIN, and then places back another mobile device for transaction. To counter this threat, R_s is transmitted by the mobile device in Transaction Request Message (message 19). R_s is generated by the SIM and is encrypted with K_c(message 7, 8), so it cannot be eavesdropped in the authentication phase. This ensures that the mobile device does not change.

D. Disclosure of relevant Information

The protocol is designed considering disclosure of information on a need to know basis. For example, *TC* is a counter that increments after each successful transaction. The record of the *TC* is kept by both, the user and the MNO. Shop POS terminal does not need to know the *TC*. In our proposed protocol, the *TC* is not exposed to POS terminal as it is a part of *TRM*. Similarly, the MNO does not need to know the shopping details of the customer. Therefore, only the total amount is transmitted to the MNO for transaction.

E. Transaction security

The transaction phase of the protocol requires maximum security. The *TRM* message is initiated by the customer rather than the shop terminal in order to satisfy the customer. The integrity of the *TRM* message is protected by the MAC so any alteration in this message is not possible. The message 19 is designed in such a way that the first half of the message containing encrypted *PI* is for shop POS terminal. POS terminal can decrypt and check the authenticity of the payment information. The remaining half of the message, containing encrypted *TRM* and corresponding MAC can neither be decrypted nor altered by the shop POS terminal. The POS terminal relays the remaining half to the MNO along with the Time Stamp. Hence, the transaction information generated by the customer is relayed to the MNO without any alteration.

In this phase, there can be a scenario where a dishonest customer has an intention of paying less than the actual amount. The customer designs a malicious *TRM* message (*TRM´*) consisting of *PI´* (an illegitimate payment information, *PI´<PI*). The dishonest customer then forms message 19 as:

$$PI = \text{Receipt Number, Total Price, Time Stamp } TS_U$$
$$TRM' = PI', R_s, \text{Transaction Counter (TC)}$$

It may be noted that the *PI* is legitimate whereas, the *TRM´* consists of amended *PI* (*PI´*). The dishonest customer forms message 19 as:

$$\text{Message 19} = E_{kc2}(PI), E_{Kc}(TRM'), MAC_{Kc1}(TRM')$$

The first half of the message, consisting of encrypted *PI*, is legitimate and the shop can verify it. However, the malicious part cannot be decrypted by the shop, so the shop cannot determine that the remaining part contains amended price information. The shop forms message 21 as *PI* is verified. The MNO executes the transaction with amended price and forms message 23 and digitally signs it. Message 23 contains the information about the amount deducted from the customer. Once this message is received by the shop terminal, the shop detects that the deducted amount is not the same as required. Hence, a dishonest customer with the intention to pay a lesser amount does not succeed in our proposed design.

F. New set of Keys for every transaction

The keys are generated from random number *R* (generated by the MNO). The *R* acts as a seed for all keys. As *R* is fresh for every transaction, therefore the keys are also new in each transaction.

G. Non-repudiation of transaction messages

The transaction result messages (message 23, message 25) are digitally signed. In case of any dispute about the payment, the MNO is to honour message 23 as it contains the MNO's digital signature. The shopping detail is also digitally signed by the shop POS terminal so the shop has to honour the prices mentioned in this message. Therefore the customer is completely secured about the transaction.

H. Securing long term secret

K_p is the long term secret between MNO and shop POS terminal. In our protocol, $K_p$is used with the least exposure (only once). The security policy of the MNO can define the update of this key after a defined interval.

VIII. Conclusion

In this paper, we have proposed a transaction protocol that provides a secure and trusted communication channel to the communication parties. The proposed protocol was based on NFC Cloud Wallet model for secure cloud-based NFC transactions. The operations performed by the vendor’s reader, an NFC enabled phone and the cloud provider (in this paper MNO) are provided and such operations are possible by the current state of the technology as most of these measures are already implemented to support other mechanisms. We considered the detailed execution of the protocol and we showed our protocol performs reliably in cloud-based NFC transaction architecture. In addition, this paper provides other related issues that are required to be explored in order to find out different possible roles, accesses and ownerships of involved parties in NFC ecosystem. The main advantage of this paper is to demonstrate another way of payment for all those people who do not have bank accounts. This way of making payments eases the process of purchasing for ordinary people as they only have to top up with their MNO without having to follow all the banking procedures.

From the business and customers' point of view, this method of payment is a micropayment, involving a very small amount for transactions (typically less than £50.00); therefore, people can only use this payment method to pay for cheap products. This has also prevented the emergence of such system, as the cost for individual transactions must be kept low which is impractical.

A. Future work

As a part of future work, a proof of concept implementation can be carried out in order to determine the reliability of the proposed protocol in terms of number of factors such as timing issues. This implementation refers to the performance domain of the proposed protocol which can be taken into the account to consider the performance of the protocol rather than its security that is discussed in this paper. The idea of the proposed protocol can also be extended to a multi-party protocol. Furthermore, other possible architectures in this area should be explored and defined in order to finalize the most reliable architecture for cloud-based NFC payment applications.

REFERENCES

[1] P. Pourghomi, and G. Ghinea "Managing NFC payments applications through cloud computing," In 7th International Conference for Internet Technology and Secured Transactions (ICITST).IEEE, pp. 772–777, December 2012.

[2] G. Madlmayr, J. Langer, J. Scharinger, "Managing an NFC ecosystem," In Proceedings of the 7th International Conference on Mobile Business, Washington, DC, USA: IEEE Computer Society, pp. 95–101, 2008.

[3] P. Pourghomi, and G. Ghinea, "Challenges of managing secure elements within the NFC ecosystem," in 7th International Conference for Internet Technology and Secured Transactions (ICITST). IEEE, pp. 720–725, December 2012.

[4] NFC Forum,"Essentials for successful NFC mobile ecosystems," 2008. www.nfcforum.org/resources/white papers/NFC Forum Mobile NFC Ecosystem White Paper.pdf

[5] M. Reveilhac and M. Pasquet, "Promising secure element alternatives for NFC technology," In: First International Workshop on Near FieldCommunication, IEEE, pp. 75 – 80. 2009.

[6] J. Eberspaecher, H. J. Voegel, C. Bettstetter, "GSM - Architecture,Protocols and Services", 3rd Ed., Wiley, New York.

[7] Y. Li, Y. Chen, T. J. Ma. "Security in GSM" www.gsmsecurity.net/papers/securityingsm.pdf

[8] "GSM System Overview" www.pcs.csie.ntu.edu.tw/course/pcs/2007/reference/04 GSM System Overview.pdf

[9] "GSM (and PCN) Security and Encryption" www.brookson.com/gsm/gsmdoc.pdf

[10] W. Chen, G. Hancke, K.Mayes, Y. Lien, Y, J.H. Chiu, "NFC mobile transactions and authentication based on GSM network" In InternationalWorkshop on Near Field Communication,IEEEComputer Society, pp. 83–89. 2010.

[11] Technical Specification Group Core Network, "Numbering, addressing and identification," 1999. www.arib.or.jp/english/html/overview/doc/STD-T63v9 30/5 Appendix/R99/23/23003-3f0.pdf

[12] M. Bellare, C. Namprempre,"Authenticated encryption: relations among notions and analysis of the generic composition paradigm," In: International

[13] Conference on Advances in Cryptology ASIACRYPT, pp. 531–545. 2000

An Analysis of Brand Selection

Kazuhiro Takeyasu†
College of Business Administration, Tokoha University,
325 Oobuchi, Fuji City, Shizuoka, 417-0801,
Japan

Yuki Higuchi
Faculty of Business Administration, Setsunan University,
17-8 Ikeda-nakamachi, Neyagawa, Osaka, 572-8508,
Japan

***Abstract*—It is often observed that consumers select upper class brand when they buy next time. Suppose that former buying data and current buying data are gathered. Also suppose that upper brand is located upper in the variable array. Then the transition matrix becomes upper triangular matrix under the supposition that former buying variables are set input and current buying variables are set output. Takeyasu et al. analyzed the brand selection and its matrix structure before. In that paper, products of one genre are analyzed. In this paper, brand selection among multiple genre and its matrix structure are analyzed. Taking an automobile for example, customer brand selection from company A to B or company A to C can be made clear utilizing above stated method. We can confirm not only the preference shift among brands but also the preference shift among companies. This enables building marketing strategy for automobile company much easier. Analyzing such structure provides useful applications. Thus, this proposed approach enables to make effective marketing plan and/or establishing new brand.**

Keywords— brand selection; matrix structure; brand position

I. Introduction

Marketing analysis is one of the "never ending themes" because there arise lots of events and new trend in the market and society. There are many themes to be investigated and the analyses may be utilized for marketing plan etc. In this paper, we focus on the brand selection by consumers.

It is often observed that consumers select upper class brand when they buy next time after they are bored to use current brand. Suppose that former buying data and current buying data are gathered. Also suppose that upper brand is located upper in the variable array. Then the transition matrix becomes upper triangular matrix under the supposition that former buying variables are set input and current buying variables are set output. The analysis of the brand selection in the same brand group is analyzed by Takeyasu et al. [6].

In this paper, we expand this scheme to products of multiple genres. For example, we consider the case of necklace. If she is accustomed to use necklace, she would buy higher priced necklace. On the other hand, she may buy bracelet or earring for her total coordination in fashion. Hearing from the retailer, both can be seen in selecting upper class brand and selecting another genre product.

Therefore, this analysis is very meaningful for the practical use, which occurs actually. If transition matrix is identified, we can make various analyses using it and s-step forecasting can be executed. Unless planners for products notice its brand position whether it is upper or lower than other products, matrix structure makes it possible to identify those by calculating consumers' activities for brand selection. Thus, this proposed approach makes it effective to execute marketing plan and/or establish new brand.

Quantitative analysis concerning brand selection has been executed by Yamanaka [5], Takahashi et al.[4]. Yamanaka[5] examined purchasing process by Markov Transition Probability with the input of advertising expense. Takahashi et al.[4] made analysis by the Brand Selection Probability model using logistics distribution. Takeyasu et al.[6] analyzed the preference shift of customer brand selection in the case of automobile.

Takeyasu et al.[7] analyzed the preference shift of customer brand selection for a single brand group. In this paper, we try to expand this scheme to products of multiple genres, and various analyses is executed. Actually, this scheme can often be seen. Such research is quite a new one.

Hereinafter, matrix structure for a single brand group is clarified for the selection of brand in section 2. Expansion to multiple brand selection is executed and analyzed in section 3. s-step forecasting is stated in section 4. Numerical calculation is executed in section 5. Application of this method is extended in section 6.

II. Brand Selection And Its Matrix Structure

A. Upper shift of Brand selection

Now, suppose that x is the most upper class brand, y is the second upper class brand, and z is the lowest class brand. Consumer's behavior of selecting brand might be $z \to y$, $y \to x$, $z \to x$ etc. $x \to z$ might be few.

Suppose that x is current buying variable, and x_b is previous buying variable. Shift to x is executed from

x_b, y_b, or z_b.

Therefore, x is stated in the following equation. a_{ij} Represents the transition probability from j-th to i-th brand.

$$x = a_{11}x_b + a_{12}y_b + a_{13}z_b$$

Similarly,

$$y = a_{22}y_b + a_{23}z_b$$

and

$$z = a_{33} z_b$$

These are re-written as follows.

$$\begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{pmatrix} a_{11} & a_{12} & a_{13} \\ 0 & a_{22} & a_{23} \\ 0 & 0 & a_{33} \end{pmatrix} \begin{pmatrix} x_b \\ y_b \\ z_b \end{pmatrix} \tag{1}$$

Set

$$\mathbf{X} = \begin{pmatrix} x \\ y \\ z \end{pmatrix},\ \mathbf{A} = \begin{pmatrix} a_{11} & a_{12} & a_{13} \\ 0 & a_{22} & a_{23} \\ 0 & 0 & a_{33} \end{pmatrix},\ \mathbf{X_b} = \begin{pmatrix} x_b \\ y_b \\ z_b \end{pmatrix}$$

Then, $\mathbf{X}$ is represented as follows.

$$\mathbf{X} = \mathbf{A}\mathbf{X_b} \tag{2}$$

Here,

$$\mathbf{X} \in \mathbf{R}^3, \mathbf{A} \in \mathbf{R}^{3\times 3}, \mathbf{X_b} \in \mathbf{R}^3$$

$\mathbf{A}$ is an upper triangular matrix.

To examine this, generating the following data, which all consist of the data in which the transition is made from lower brand to upper brand,

$$\mathbf{X^i} = \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix} \quad \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix} \quad \cdots \quad \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix} \tag{3}$$

$$\mathbf{X_b^i} = \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix} \quad \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix} \quad \cdots \quad \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix} \tag{4}$$

$$i = 1, \qquad 2 \qquad \cdots \qquad N$$

Parameter can be estimated by using least square method.

Suppose

$$\mathbf{X}^i = \mathbf{A}\mathbf{X_b}^i + \boldsymbol{\varepsilon}^i \tag{5}$$

Where

$$\varepsilon^i = \begin{pmatrix} \varepsilon_1^i \\ \varepsilon_2^i \\ \varepsilon_3^i \end{pmatrix} \qquad i = 1, 2, \cdots, N$$

and minimize following J

$$J = \sum_{i=1}^{N} \boldsymbol{\varepsilon}^{iT} \boldsymbol{\varepsilon}^i \to Min \tag{6}$$

$\hat{\mathbf{A}}$ Which an estimated value is of $\mathbf{A}$ is obtained as follows.

$$\hat{\mathbf{A}} = \left(\sum_{i=1}^{N} \mathbf{X}^i \mathbf{X_b}^{iT} \right) \left(\sum_{i=1}^{N} \mathbf{X_b}^i \mathbf{X_b}^{iT} \right)^{-1} \tag{7}$$

In the data group which all consist of the data in which the transition is made from lower brand to upper brand, estimated value $\hat{\mathbf{A}}$ should be upper triangular matrix. If the following data which shift to lower brand are added only a few in equation (3) and (4),

$$\mathbf{X}^i = \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix},\quad \mathbf{X_b}^i = \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix}$$

$\hat{\mathbf{A}}$ would contain minute items in the lower part of triangle.

B. Sorting brand ranking by re-arranging row

In a general data, variables may not be in order as x, y, z . In that case, large and small value lie scattered in $\hat{\mathbf{A}}$. But re-arranging this, we can set in order by shifting row. The large value parts are gathered in the upper triangular matrix, and the small value parts are gathered in the lower triangular matrix.

$$\begin{array}{cc} & \hat{\mathbf{A}} \\ \begin{pmatrix} x \\ y \\ z \end{pmatrix} & \begin{pmatrix} \bigcirc & \bigcirc & \bigcirc \\ \varepsilon & \bigcirc & \bigcirc \\ \varepsilon & \varepsilon & \bigcirc \end{pmatrix} \end{array} \underset{\Longleftarrow}{\text{Shifting row}} \begin{array}{cc} & \hat{\mathbf{A}} \\ \begin{pmatrix} z \\ x \\ y \end{pmatrix} & \begin{pmatrix} \varepsilon & \varepsilon & \bigcirc \\ \bigcirc & \bigcirc & \bigcirc \\ \varepsilon & \bigcirc & \bigcirc \end{pmatrix} \end{array} \tag{8}$$

C. Matrix structure under the case skipping intermediate class brand is skipped

It is often observed that some consumers select the most upper class brand from the most lower class brand and skip selecting the intermediate class brand.

We suppose v, w, x, y, z brands (suppose they are laid from upper position to lower position as $v > w > x > y > z$). In the above case, selection shifts would be:

$$v \leftarrow z$$

$$v \leftarrow y$$

Suppose they do not shift to y, x, w from z, to x, w from y, and to w from x, then Matrix structure would be as follows.

$$\begin{pmatrix} v \\ w \\ x \\ y \\ z \end{pmatrix} = \begin{pmatrix} a_{11} & a_{12} & a_{13} & a_{14} & a_{15} \\ 0 & a_{22} & 0 & 0 & 0 \\ 0 & 0 & a_{33} & 0 & 0 \\ 0 & 0 & 0 & a_{44} & 0 \\ 0 & 0 & 0 & 0 & a_{55} \end{pmatrix} \begin{pmatrix} v_b \\ w_b \\ x_b \\ y_b \\ z_b \end{pmatrix} \quad (9)$$

We confirm this by the numerical example in section 4.

III. Expansion Of The Model To Multiple Genre Products

Expanding Eq.(2) to multiple genre products, we obtain the following equations. First of all, we state the generalized model of Eq.(2).

$$\mathbf{X} = \mathbf{A}\mathbf{X_b} \quad (10)$$

Where

$$\mathbf{X} = \begin{pmatrix} x^1 \\ x^2 \\ \vdots \\ x^p \end{pmatrix}, \ \mathbf{X_b} = \begin{pmatrix} x_b^1 \\ x_b^2 \\ \vdots \\ x_b^p \end{pmatrix} \quad (11)$$

$$\mathbf{A} = \begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1p} \\ a_{21} & a_{22} & \cdots & a_{2p} \\ \vdots & \vdots & & \vdots \\ a_{p1} & a_{p2} & \cdots & a_{pp} \end{pmatrix} \quad (12)$$

Here

$$\mathbf{X} \in \mathbf{R}^p, \mathbf{A} \in \mathbf{R}^{p\times p}, \mathbf{X_b} \in \mathbf{R}^p$$

If the brand selection is executed towards upper class, then $\mathbf{A}$ becomes as follows.

$$\mathbf{A} = \begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1p} \\ 0 & a_{22} & \cdots & a_{2p} \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \cdots & a_{pp} \end{pmatrix} \quad (13)$$

Expanding above equations to products of 3 genres, we obtain the following equations.

$$\begin{pmatrix} \mathbf{X} \\ \mathbf{Y} \\ \mathbf{Z} \end{pmatrix} = \begin{pmatrix} \mathbf{A}^{11} & \mathbf{A}^{12} & \mathbf{A}^{13} \\ \mathbf{A}^{21} & \mathbf{A}^{22} & \mathbf{A}^{23} \\ \mathbf{A}^{31} & \mathbf{A}^{32} & \mathbf{A}^{33} \end{pmatrix} \begin{pmatrix} \mathbf{X_b} \\ \mathbf{Y_b} \\ \mathbf{Z_b} \end{pmatrix} \quad (14)$$

Where

$$\mathbf{X} = \begin{pmatrix} x^1 \\ x^2 \\ \vdots \\ x^p \end{pmatrix}, \ \mathbf{X_b} = \begin{pmatrix} x_b^1 \\ x_b^2 \\ \vdots \\ x_b^p \end{pmatrix}, \ \mathbf{Y} = \begin{pmatrix} y^1 \\ y^2 \\ \vdots \\ y^q \end{pmatrix},$$

$$\mathbf{Y_b} = \begin{pmatrix} y_b^1 \\ y_b^2 \\ \vdots \\ y_b^q \end{pmatrix}, \ \mathbf{Z} = \begin{pmatrix} z^1 \\ z^2 \\ \vdots \\ z^r \end{pmatrix}, \ \mathbf{Z_b} = \begin{pmatrix} z_b^1 \\ z_b^2 \\ \vdots \\ z_b^r \end{pmatrix} \quad (15)$$

$$\mathbf{A}^{11} = \begin{pmatrix} a_{11}^{11} & a_{12}^{11} & \cdots & a_{1p}^{11} \\ a_{21}^{11} & a_{22}^{11} & \cdots & a_{2p}^{11} \\ \vdots & \vdots & & \vdots \\ a_{p1}^{11} & a_{p2}^{11} & \cdots & a_{pp}^{11} \end{pmatrix},$$

$$\mathbf{A}^{12} = \begin{pmatrix} a_{11}^{12} & a_{12}^{12} & \cdots & a_{1q}^{12} \\ a_{21}^{12} & a_{22}^{12} & \cdots & a_{2q}^{12} \\ \vdots & \vdots & & \vdots \\ a_{p1}^{12} & a_{p2}^{12} & \cdots & a_{pq}^{12} \end{pmatrix},$$

$$\mathbf{A}^{13} = \begin{pmatrix} a_{11}^{13} & a_{12}^{13} & \cdots & a_{1r}^{13} \\ a_{21}^{13} & a_{22}^{13} & \cdots & a_{2r}^{13} \\ \vdots & \vdots & & \vdots \\ a_{p1}^{13} & a_{p2}^{13} & \cdots & a_{pr}^{13} \end{pmatrix},$$

$$\mathbf{A}^{21} = \begin{pmatrix} a_{11}^{21} & a_{12}^{21} & \cdots & a_{1p}^{21} \\ a_{21}^{21} & a_{22}^{21} & \cdots & a_{2p}^{21} \\ \vdots & \vdots & & \vdots \\ a_{q1}^{21} & a_{q2}^{21} & \cdots & a_{qp}^{21} \end{pmatrix}, \quad (16)$$

$$\mathbf{A}^{22} = \begin{pmatrix} a_{11}^{22} & a_{12}^{22} & \cdots & a_{1q}^{22} \\ a_{21}^{22} & a_{22}^{22} & \cdots & a_{2q}^{22} \\ \vdots & \vdots & & \vdots \\ a_{q1}^{22} & a_{q2}^{22} & \cdots & a_{qq}^{22} \end{pmatrix},$$

$$\mathbf{A}^{23} = \begin{pmatrix} a_{11}^{23} & a_{12}^{23} & \cdots & a_{1r}^{23} \\ a_{21}^{23} & a_{22}^{23} & \cdots & a_{2r}^{23} \\ \vdots & \vdots & & \vdots \\ a_{q1}^{23} & a_{q2}^{23} & \cdots & a_{qr}^{23} \end{pmatrix},$$

$$\mathbf{A}^{31}=\begin{pmatrix} a_{11}^{31} & a_{12}^{31} & \cdots & a_{1p}^{31} \\ a_{21}^{31} & a_{22}^{31} & \cdots & a_{2p}^{31} \\ \vdots & \vdots & & \vdots \\ a_{r1}^{31} & a_{r2}^{31} & \cdots & a_{rp}^{31} \end{pmatrix},$$

$$\mathbf{A}^{32}=\begin{pmatrix} a_{11}^{32} & a_{12}^{32} & \cdots & a_{1q}^{32} \\ a_{21}^{32} & a_{22}^{32} & \cdots & a_{2q}^{32} \\ \vdots & \vdots & & \vdots \\ a_{r1}^{32} & a_{r2}^{32} & \cdots & a_{rq}^{32} \end{pmatrix},$$

$$\mathbf{A}^{33}=\begin{pmatrix} a_{11}^{33} & a_{12}^{33} & \cdots & a_{1r}^{33} \\ a_{21}^{33} & a_{22}^{33} & \cdots & a_{2r}^{33} \\ \vdots & \vdots & & \vdots \\ a_{r1}^{33} & a_{r2}^{33} & \cdots & a_{rr}^{33} \end{pmatrix}$$

$\mathbf{X}\in\mathbf{R}^{p},\mathbf{X_b}\in\mathbf{R}^{p},\mathbf{Y}\in\mathbf{R}^{q},\mathbf{Y_b}\in\mathbf{R}^{q},\mathbf{Z}\in\mathbf{R}^{r},$
$\mathbf{Z_b}\in\mathbf{R}^{r},\mathbf{A}^{11}\in\mathbf{R}^{p\times p},\mathbf{A}^{12}\in\mathbf{R}^{p\times q},\mathbf{A}^{13}\in\mathbf{R}^{p\times r},$
$\mathbf{A}^{21}\in\mathbf{R}^{q\times p},\mathbf{A}^{22}\in\mathbf{R}^{q\times q},\mathbf{A}^{23}\in\mathbf{R}^{q\times r},$
$\mathbf{A}^{31}\in\mathbf{R}^{r\times p},\mathbf{A}^{32}\in\mathbf{R}^{r\times q},\mathbf{A}^{33}\in\mathbf{R}^{r\times r}$

Re-writing Eq.(14) as :

$$\mathbf{W}=\mathbf{A}\mathbf{W_b} \tag{17}$$

Then, the transition matrix $\mathbf{A}$ is derived as follows in the same way with Eq.(7).

$$\hat{\mathbf{A}}=\left(\sum_{i=1}^{N}\mathbf{W}^{i}\mathbf{W}_{\mathbf{b}}^{iT}\right)\left(\sum_{i=1}^{N}\mathbf{W}_{\mathbf{b}}^{i}\mathbf{W}_{\mathbf{b}}^{iT}\right)^{-1} \tag{18}$$

Here,

$$\mathbf{W}=\begin{pmatrix}\mathbf{X}\\ \mathbf{Y}\\ \mathbf{Z}\end{pmatrix},\ \mathbf{W_b}=\begin{pmatrix}\mathbf{X_b}\\ \mathbf{Y_b}\\ \mathbf{Z_b}\end{pmatrix} \tag{19}$$

$$\mathbf{A}=\begin{pmatrix}\mathbf{A}^{11} & \mathbf{A}^{12} & \mathbf{A}^{13}\\ \mathbf{A}^{21} & \mathbf{A}^{22} & \mathbf{A}^{23}\\ \mathbf{A}^{31} & \mathbf{A}^{32} & \mathbf{A}^{33}\end{pmatrix} \tag{20}$$

$$\mathbf{W}^{i}=\mathbf{A}\mathbf{W}_{\mathbf{b}}^{i}+\boldsymbol{\varepsilon}^{i}\quad i=1,2,\cdots,N \tag{21}$$

$$\boldsymbol{\varepsilon}^{i}=\begin{pmatrix}\varepsilon_{1}^{i}\\ \vdots\\ \varepsilon_{p}^{i}\\ \varepsilon_{p+1}^{i}\\ \vdots\\ \varepsilon_{p+q}^{i}\\ \varepsilon_{p+q+1}^{i}\\ \vdots\\ \varepsilon_{p+q+r}^{i}\end{pmatrix}\quad i=1,2,\cdots,N \tag{22}$$

If the brand selection is executed towards upper class brand in the same genre, the transition matrix, for example

$\mathbf{A}_{11}$, $\mathbf{A}_{22}$, $\mathbf{A}_{33}$, become an upper triangular matrix as can be seen in 2. Suppose $\mathbf{X}$ as bracelet, $\mathbf{Y}$ as earring and $\mathbf{Z}$ as necklace. If we only see $\mathbf{Z}$, we can examine whether there is an upper brand shift in $\mathbf{A}_{33}$. But there is a case that brand selection is executed towards other genre products. There occurs brand selection shift from a certain brand level of $\mathbf{Z}$ to a certain brand level of $\mathbf{X}$ or $\mathbf{Y}$. For example, suppose there are five levels in each $\mathbf{X}$, $\mathbf{Y}$, $\mathbf{Z}$ and their levels include from bottom to top brand level. In that case, if there is a brand selection shift from the middle brand level in $\mathbf{Z}$ to another genre product, we can obtain interesting result by examining how the brand selection shift is executed toward the same level or upper level of another genre product. If we can see the trend of brand selection shift, we can foresee the brand selection shift towards another genre brand. Retailer can utilize the result of this to make effective marketing plan. We confirm this by the simple numerical example in 5.

Next, we examine the case in brand groups. Matrices are composed by Block Matrix.

IV. *S*-STEP FORECASTING

Now, we see Eq.(14) in time series. Set $\mathbf{X}$, $\mathbf{Y}$, $\mathbf{Z}$ at time n as :

$$\mathbf{X_n}=\begin{pmatrix}x_{1}^{n}\\ x_{2}^{n}\\ \vdots\\ x_{p}^{n}\end{pmatrix},\mathbf{Y_n}=\begin{pmatrix}y_{1}^{n}\\ y_{2}^{n}\\ \vdots\\ y_{q}^{n}\end{pmatrix},\mathbf{Z_n}=\begin{pmatrix}z_{1}^{n}\\ z_{2}^{n}\\ \vdots\\ z_{r}^{n}\end{pmatrix} \tag{23}$$

Then, Eq.(14) can be re-stated as :

$$\begin{pmatrix}\mathbf{X_n}\\ \mathbf{Y_n}\\ \mathbf{Z_n}\end{pmatrix}=\begin{pmatrix}\mathbf{A}_{11} & \mathbf{A}_{12} & \mathbf{A}_{13}\\ \mathbf{A}_{21} & \mathbf{A}_{22} & \mathbf{A}_{23}\\ \mathbf{A}_{31} & \mathbf{A}_{32} & \mathbf{A}_{33}\end{pmatrix}\begin{pmatrix}\mathbf{X_{n-1}}\\ \mathbf{Y_{n-1}}\\ \mathbf{Z_{n-1}}\end{pmatrix} \tag{24}$$

Where suffix is written in the lower part of right hand side because there arises a multiplier in the equation of forecasting.

s -step forecasting is executed by the following equation.

$$\begin{pmatrix} \mathbf{X_{n+s}} \\ \mathbf{Y_{n+s}} \\ \mathbf{Z_{n+s}} \end{pmatrix} = \begin{pmatrix} \mathbf{A}_{11} & \mathbf{A}_{12} & \mathbf{A}_{13} \\ \mathbf{A}_{21} & \mathbf{A}_{22} & \mathbf{A}_{23} \\ \mathbf{A}_{31} & \mathbf{A}_{32} & \mathbf{A}_{33} \end{pmatrix}^{s} \begin{pmatrix} \mathbf{X_n} \\ \mathbf{Y_n} \\ \mathbf{Z_n} \end{pmatrix} \tag{25}$$

V. NUMERICAL EXAMPLE

We consider the case of $p = q = r = 4$ in 3. Suppose there are following customer preference shifts.

From the lower level of $\mathbf{Z}$ to middle or upper level of $\mathbf{Z}$

From the lower level of $\mathbf{Z}$ to lower, middle or upper level of $\mathbf{Y}$

From the middle level of $\mathbf{Z}$ to middle or upper level of $\mathbf{Y}$

From the upper level of $\mathbf{Z}$ to upper level of $\mathbf{Y}$

From the lower level of $\mathbf{Z}$ to lower, middle or upper level of $\mathbf{X}$

From the middle level of $\mathbf{Z}$ to middle or upper level of $\mathbf{X}$

From the upper level of $\mathbf{Z}$ to upper level of $\mathbf{X}$

And also suppose that there are preference shifts which stay at the same level in $\mathbf{Z}$. For simplicity, preference shift of $\mathbf{X}, \mathbf{Y}$ stays at the same level within $\mathbf{X}$ and $\mathbf{Y}$. In these cases, we can assume that each block matrix in $\mathbf{A}$ of Eq.(21) becomes as follows.

$$\mathbf{A}_{11}, \mathbf{A}_{22}: \text{ Diagonal matrix}$$

$$\mathbf{A}_{12}, \mathbf{A}_{21}, \mathbf{A}_{31}, \mathbf{A}_{32}: \ \mathbf{0}$$

$$\mathbf{A}_{13}, \mathbf{A}_{23}, \mathbf{A}_{33}: \text{ Upper triangular matrix}$$

Now we suppose customer preference shifts as follows.

1.	Jump from 4th to 3rd rank of $\mathbf{Z}$	:	2 events
2.	Jump from 4th to 2nd rank of $\mathbf{Z}$	:	1 event
3.	Jump from 2nd to 1st rank of $\mathbf{Z}$	:	2 events
4.	Stay at 4th rank of $\mathbf{Z}$	:	3 events
5.	Stay at 3rd rank of $\mathbf{Z}$	:	4 events
6.	Stay at 2nd rank of $\mathbf{Z}$	:	4 events
7.	Stay at 1st rank of $\mathbf{Z}$	:	2 events
8.	Jump from 4th rank of $\mathbf{Z}$ to 3rd rank of $\mathbf{Y}$	:	1 event
9.	Jump from 4th rank of $\mathbf{Z}$ to 2nd rank of $\mathbf{Y}$	:	1 event
10.	Jump from 3rd rank of $\mathbf{Z}$ to 1st rank of $\mathbf{Y}$	:	1 event
11.	Stay at 4th rank of $\mathbf{Y}$	:	2 events
12.	Stay at 3rd rank of $\mathbf{Y}$	:	3 events
13.	Stay at 2nd rank of $\mathbf{Y}$	:	1 event
14.	Stay at 1st rank of $\mathbf{Y}$	:	2 events
15.	Jump from 4th rank of $\mathbf{Z}$ to 3rd rank of $\mathbf{X}$	:	2 events
16.	Jump from 3rd rank of $\mathbf{Z}$ to 1st rank of $\mathbf{X}$	:	1 events
17.	Stay at 4th rank of $\mathbf{X}$	:	3 events
18.	Stay at 3rd rank of $\mathbf{X}$	:	2 events
19.	Stay at 2nd rank of $\mathbf{X}$	:	3 events
20.	Stay at 1st rank of $\mathbf{X}$	:	1 event
21.	Jump from 4th rank of $\mathbf{Z}$ to 4th rank of $\mathbf{Y}$	:	2 events
22.	Jump from 4th rank of $\mathbf{Z}$ to 2nd rank of $\mathbf{X}$	:	2 events
23.	Jump from 2nd rank of $\mathbf{Z}$ to 2nd rank of $\mathbf{Y}$	:	2 events
24.	Jump from 1st rank of $\mathbf{Z}$ to 1st rank of $\mathbf{Y}$	:	1 event
25.	Jump from 1st rank of $\mathbf{Z}$ to 1st rank of $\mathbf{X}$	:	1 event
26.	Jump from 3rd rank of $\mathbf{Z}$ to 3rd rank of $\mathbf{Y}$	:	2 events
27.	Jump from 3rd rank of $\mathbf{Z}$ to 3rd rank of $\mathbf{X}$	:	1 event
28.	Jump from 2nd rank of $\mathbf{Z}$ to 2nd rank of $\mathbf{X}$	:	2 events
29.	Jump from 2nd rank of $\mathbf{Z}$ to 1st rank of $\mathbf{X}$	:	1 event

Then, the vector $\mathbf{W}, \mathbf{W_b}$ for case 1-2, for example, are expressed as follows.

$$1. \quad \mathbf{W} = \begin{pmatrix} 0 \\ 0 \\ 0 \\ \vdots \\ 0 \\ 0 \\ 0 \\ 0 \\ \vdots \\ 0 \\ 0 \\ 0 \\ 1 \\ 0 \end{pmatrix} \quad \cdots \quad \mathbf{W_b} = \begin{pmatrix} 0 \\ 0 \\ 0 \\ \vdots \\ 0 \\ 0 \\ 0 \\ 0 \\ \vdots \\ 0 \\ 0 \\ 0 \\ 0 \\ 1 \end{pmatrix} \quad 2. \quad \mathbf{W} = \begin{pmatrix} 0 \\ 0 \\ 0 \\ \vdots \\ 0 \\ 0 \\ 0 \\ 0 \\ \vdots \\ 0 \\ 0 \\ 1 \\ 0 \\ 0 \end{pmatrix} \quad \mathbf{W_b} = \begin{pmatrix} 0 \\ 0 \\ 0 \\ \vdots \\ 0 \\ 0 \\ 0 \\ 0 \\ \vdots \\ 0 \\ 0 \\ 0 \\ 0 \\ 1 \end{pmatrix}$$

Substituting these to Eq.(18), we can obtain Eq.(26).

$$\mathbf{A}=\left(\begin{array}{cccc:cccc:cccc}
1&0&0&0&0&0&0&0&1&1&1&0\\
0&3&0&0&0&0&0&0&0&2&0&0\\
0&0&2&0&0&0&0&0&0&0&1&2\\
0&0&0&3&0&0&0&0&0&0&0&2\\
\hdashline
0&0&0&0&2&0&0&0&1&0&1&0\\
0&0&0&0&0&1&0&0&0&2&0&1\\
0&0&0&0&0&0&3&0&0&0&2&1\\
0&0&0&0&0&0&0&2&0&0&0&2\\
\hdashline
0&0&0&0&0&0&0&0&2&2&0&0\\
0&0&0&0&0&0&0&0&0&4&0&1\\
0&0&0&0&0&0&0&0&0&0&4&2\\
0&0&0&0&0&0&0&0&0&0&0&3
\end{array}\right)\times
\left(\begin{array}{cccc:cccc:cccc}
1&0&0&0&0&0&0&0&0&0&0&0\\
0&3&0&0&0&0&0&0&0&0&0&0\\
0&0&2&0&0&0&0&0&0&0&0&0\\
0&0&0&3&0&0&0&0&0&0&0&0\\
\hdashline
0&0&0&0&2&0&0&0&0&0&0&0\\
0&0&0&0&0&1&0&0&0&0&0&0\\
0&0&0&0&0&0&3&0&0&0&0&0\\
0&0&0&0&0&0&0&2&0&0&0&0\\
\hdashline
0&0&0&0&0&0&0&0&4&0&0&0\\
0&0&0&0&0&0&0&0&0&11&0&0\\
0&0&0&0&0&0&0&0&0&0&9&0\\
0&0&0&0&0&0&0&0&0&0&0&14
\end{array}\right)^{-1} \qquad (26)$$

$$=\left(\begin{array}{cccc:cccc:cccc}
1&0&0&0&0&0&0&0&\frac{1}{4}&\frac{1}{11}&\frac{1}{9}&0\\
0&1&0&0&0&0&0&0&0&\frac{2}{11}&0&0\\
0&0&1&0&0&0&0&0&0&0&\frac{1}{9}&\frac{1}{7}\\
0&0&0&1&0&0&0&0&0&0&0&\frac{1}{7}\\
\hdashline
0&0&0&0&1&0&0&0&\frac{1}{4}&0&\frac{1}{9}&0\\
0&0&0&0&0&1&0&0&0&\frac{2}{11}&0&\frac{1}{14}\\
0&0&0&0&0&0&1&0&0&0&\frac{2}{9}&\frac{1}{14}\\
0&0&0&0&0&0&0&1&0&0&0&\frac{1}{7}\\
\hdashline
0&0&0&0&0&0&0&0&\frac{1}{2}&\frac{2}{11}&0&0\\
0&0&0&0&0&0&0&0&0&\frac{4}{11}&0&\frac{1}{14}\\
0&0&0&0&0&0&0&0&0&0&\frac{4}{9}&\frac{1}{7}\\
0&0&0&0&0&0&0&0&0&0&0&\frac{3}{14}
\end{array}\right)$$

Watching this, we can confirm following features as stated before. $\mathbf{A}$ of Eq.(21) became as follows.

$$\mathbf{A}_{11},\mathbf{A}_{22}:\ \text{Diagonal matrix}$$

$$\mathbf{A}_{12},\mathbf{A}_{21},\mathbf{A}_{31},\mathbf{A}_{32}:\ \mathbf{0}$$

$$\mathbf{A}_{13},\mathbf{A}_{23},\mathbf{A}_{33}:\ \text{Upper triangular matrix}$$

Taking an automobile for example, customer brand selection from company A to B or company A to C can be made clear utilizing above stated method. We can confirm not only the preference shift among brands but also the preference shift among companies. This enables building marketing strategy for automobile company much easier.

VI. Conclusion

Consumers often buy higher grade brand products as they are accustomed or bored to use current brand products they have.

In this paper, matrix structure was clarified when brand selection was executed toward higher grade brand. Expanding brand selection from single brand group to multiple genre brand group, we could make much more exquisite and multi-dimensional analysis. And formulation of extension to the brand groups was executed by using Block Matrix. s -step forecast model was also formulated. In numerical example, matrix structure's hypothesis was verified concerning brand selection among multiple brands. If we can see the trend of brand selection shift, we can foresee the brand selection shift towards another genre brand. Retailer can utilize the result of this to make effective marketing plan. Such research as questionnaire investigation of consumers' activity in automobile purchasing should be executed in the near future to verify obtained results.

REFERENCES

[1] Aker,D.A, Management Brands Equity, Simon & Schuster, USA,1991.

[2] Katahira,H., Marketing Science (In Japanese), Tokyo University Press, 1987.

[3] Katahira,H.,Y,Sugita., "Current movement of Marketing Science" (In Japanese), Operations Research, 1994; 14: 178-188

[4] Takahashi,Y., T.Takahashi, "Building Brand Selection Model Considering Consumers Royalty to Brand"(In Japanese), Japan Industrial Management Association 2002; 53(5): 342-347

[5] Yamanaka,H., "Quantitative Research Concerning Advertising and Brand Shift" (In Japanese), Marketing Science, Chikra-Shobo Publishing, 1982.

[6] Takeyasu,K., Y.Higuchi, "Analysis of the Preference Shift of Customer Brand Selection among Multiple Genres of Automobile and Its Matrix Structure", Journal of Communication and Computer, 20012; 9(7): 744-751

[7] Takeyasu,K., Y.Higuchi, "Analysis of the Preference Shift of Customer Brand Selection", International Journal of Computational Science, 2007; 1(4): 371-394

Protein PocketViewer: A Web-Service Based Interface for Protein Pocket Extraction and Visualization

Xiaoyu Zhang
Department of Computer Science& Information Systems
California State University San Marcos
San Marcos, U.S.A.

Martin Gordon
Department of Computer Science& Information Systems
California State University San Marcos
San Marcos, U.S.A.

***Abstract*—One important problem in bioinformatics is to study pockets or tunnels within the protein structure. These pocket or tunnel regions are significant because they indicate areas of ligand binding or enzymatic reactions, and tunnels are often solvent ion conductance areas. The Protein Pocket Viewer (PPV) is a web interface that allows the user to extract and visualize the protein pockets in a browser, based on the algorithm in** [1]**. The PPV packaged the pocket extraction executable as a web service, and made it accessible to all users with the Internet access and a modern java enabled browser. The PPV employed the Model2design pattern, which led to a loosely coupled implementation that is more robust and easier to maintain. It consists of a client web interface for user inputs and visualization, a middle-layer for controlling the flow, and the backend web services performing the actual CPU-intensive computation. The PPV web client consists of multiple window regions, with each region providing differing views of the protein, pockets and related information. For a more responsive user experience, the PPV web client employs AJAX for asynchronous execution of long running tasks, like protein pocket extraction.**

Keywords—protein structure; pockets; Model 2; AJAX; web service; visualization;

I. INTRODUCTION

Bioinformatics applies computational tools to study problems in molecular biology. Structures are critical for thefunctions of proteins. One important bioinformatics problem is to determine pockets or tunnels within the protein structure. These pocket or tunnel regions are significant because they indicate areas of ligand binding or enzymatic reactions [2], and tunnels are often solvent ion conductance areas [3]. Such computation can be data dense and computationally intensive due to the volume of data processing involved for a single protein and the number of proteins in the database. So the computations are more efficiently performed on powerful computational servers. Visualization tools are also important for bioinformatics research. A visualization tool would help the user to better comprehend and quickly consume data of the computed protein pockets. The visualization should be available for the user on the less powerful client computers, most conveniently in a web browser without installing any special software.

In this paper, we developed a web service [4] based visualization interface to the pocket extraction algorithm described in [1].The algorithm employs a two-step level set marching algorithm (Fig 1). The first step of the level set marching algorithm marches outward from the protein surface to some distance equal to a given threshold. At completion of the outward marching step, an outer surface is obtained with all indentations on original surface filled. The second step of the algorithm marches backward from the outer surface back toward the protein for the same distance. The second marching step cannot infiltratethe protein surface, or reach depressions and tunnels on the surface. The unreachable regions outside the protein surface are considered as pockets. The bounding envelops of the pockets are then extracted using standard level-set methods.

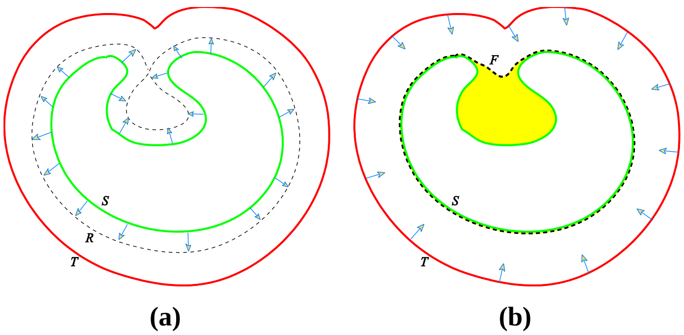
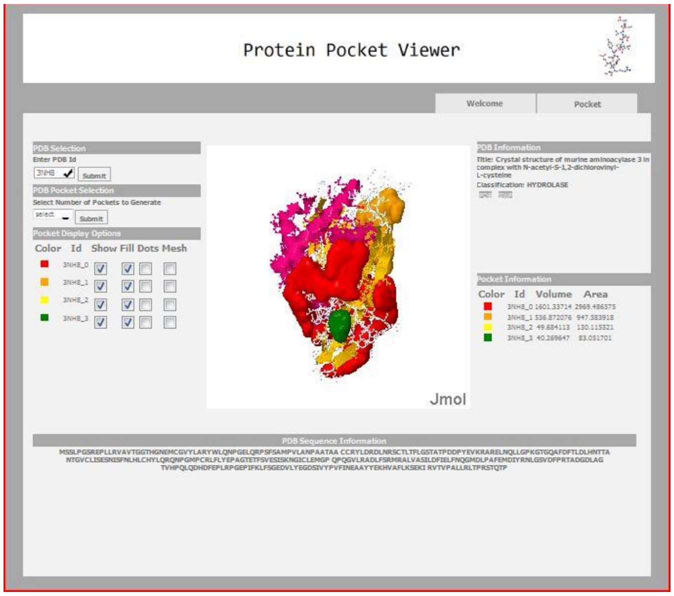


Fig. 1. The two-step level set marching algorithm for pocket extraction. (a) Outward marching from the original surface S to an outer surface T; (b) Backward marching from T to uncover the pocket as the shaded region

The visualization interface, Protein Pocket Viewer (PPV) at *http://ppv.cs.csusm.edu:8080/PPVClient/PPV.jsp*,allows users to display and manipulate data related to protein pockets in ajava-enabled browser.The web interface consists of multiple window regions, with each region providing differing views of protein pocket related data (Fig 2). The display includes both metadata about the protein and associated pockets and the three dimensional rendering of the protein and pockets. The 3D visualization of the pockets is displayed in the central rendering region. The 3D rendering can be maneuveredby the user to view of all surfaces of the protein and pockets. The 3D display can be controlled such that individual pocket can be shown or hidden, and rendered in different styles, such as a filled, dot or mesh surface. The protein sequence informationis displayed as text in the sequence region. The pocket information region displays pocket metadata such as a pocket ID, pocket surface area, and pocket volume for each pocket identified using the two-step level setmarching algorithm. The web interface also includes a protein information region that provides additional information on the protein containing the selected pockets.



3NH8 ✓

Color	Id	Show	Fill	Dots	Mesh
🟥	3NH8_0	☑	☑	☐	☐
🟧	3NH8_1	☑	☑	☐	☐
🟨	3NH8_2	☑	☑	☐	☐
🟩	3NH8_3	☑	☑	☐	☐

Title: Crystal structure of murine aminoacylase 3 in complex with N-acetyl-S-1,2-dichlorovinyl-L-cysteine

Classification: HYDROLASE

Color	Id	Volume	Area
🟥	3NH8_0	1601.33714	2969.486575
🟧	3NH8_1	536.872076	947.583918
🟨	3NH8_2	49.684113	130.115521
🟩	3NH8_3	40.269647	83.051701

MSSLPGSREPLLRVAVTGGTHGNEMCGVYLARYWLQNPGELQRPSFSAMPVLANPAATAA CCRYLDRDLNRSCTLTFLGSTATPDDPYEVKRARELNQLLGPKGTGQAFDFTLDLHNTTA
NTGVCLISESNISFNLHLCHYLQRQNPGMPCRLFLYEPAGTETFSVESISKNGICLEMGP QPQGVLRADLFSRMRALVASILDFIELFNQGMDLPAFEMDIYRNLGSVDFPRTADGDLAG
TVHPQLQDHDFEPLRPGEPIFKLFSGEDVLYEGDSIVYPVFINEAAYYEKHVAFLKSEKI RVTVPALLRLTPRSTQTP

Fig. 2. Protein Pocket Viewer interface.

The PPV allows the user to visualize the protein pockets and their relationship to the protein, and is accessible to all users with the Internet access and a modern java enabled browser. The PPV implemented the web client, a web service wrapper for invoking pocket generation, Jmol[5]viewer integration, and cache management for protein and pocket files. It has some unique features, such as

- On-the-fly generation and visualization of pockets given a valid Protein Data Bank (PDB)[6]ID or a well-formed PDB file using the two-step level setmarching algorithm.
- Caching management capability where expensive pocket generation results are cached for the later requests.
- Asynchronous transactions via AJAX [7]relieving the user from waiting for page refresh during long running tasks, such as pocket generation and PDB metadata retrieval.

In the rest of the paper, section 2 discusses background concepts used by the PPV and related work. Section 3 presents the design of the PPV and implementation details. We then conclude and discuss some future directions in section 4.

II. Background And Related Work

A. Background

The Protein Pocket Viewer employs the implementation for protein pocket extraction algorithm described in [1], which was implemented as a C++ program running on a computational server.

The Protein Data Bank (PDB) is the single worldwide repository of information about the 3D structures of large biological molecules[6], managed by the Research Collaboratory for Structural Bioinformatics (RCSB). Every protein in the PDB has a unique ID and its structure is available in PDB file format. The pocket extraction program requires a valid PDB ID or a well-formed PDB file as input.

A web service [4]is used to initiate the pocket extraction implementation. This web service provides the abstraction of the pocket extraction program and enables a bridge between the client Java classes to the pocket extraction service implemented in C++. Protein metadata, PDB file, and pocket files transfer between the server and the client via the web service interface.A web service is an implementation of the SOA design approach employing XML [8], XSD [8], and SOAP [9] standards based technologies. The Service Oriented Architecture (SOA) is a design consideration where its solution is distributed, loosely coupled, standards based, reusable, and stateless.The Web Service Description Language (WSDL)[10] provides a way for a web service provider to describe the public interface of the web service.

The 3D rendering of the proteins and their respective pockets in a browser was performed with Jmol[5], which is an open source Java viewer for 3D chemical structures. Jmol includes features for element display selection, scheme selection, element color assignment, surface display selection, and measurements. Element display selection allows the user to select the elements within the chemical structure to display. Scheme selection allows the user to select schemes such as CPK space-fill, ball and stick, sticks, wireframe, and cartoon for the chemical structure. Surface display selection allows the user to select how the surface is displayed.

The PPV utilizes Java server pages [11], cascading stylesheets [12], and JavaScript XHTML DOM scripting [13] for designing and displaying the website content. JavaScript [14], AJAX [7], and Java servlets [11] provide the flow and navigation capabilities within the viewer.

The Protein Pocket Viewer employs the Model 2 design pattern[15], which is a variant of the Model-View-Controller (MVC) design pattern. Model2 extends the Model-View-Controller design pattern for use with web applications. Model2 can employ Java server pages and cascading stylesheets to implement the view of the pages within the website. Java server pages (JSP) enable dynamic content within HTML pages. Cascading stylesheets are used to describe the look and feel of the content to display. The Java server pages and servlets provide the controller responsibility of the implementation. Behind the scenes, some Java classes provide the functionality that enable retrieval of the protein, the pockets, and the associated metadata.

There are many advantages to using the Model2 design patterns[15]. The main advantage of the design pattern is to separate the different levels of concerns within the implementation. Keeping the levels of concerns separate helps facilitate cleaner, and more loosely coupled implementation, which lowers the dependencies between different components. It allows the developer to break the implementation into smaller more manageable components. Each component can be tested separately to ensure its correctness. This makes the implementation more maintainable.

Asynchronous JavaScript and XML (AJAX) [7]provides the ability to initiate functionality asynchronously. AJAX frees up the user from busy waiting when the request is being processed at the server end. In an interactive program like a web interface, the user wants to continue interacting with the interface while waiting for a long-time processing or data retrieval. For example, a graphical user interface that performs a long running calculation while allowing the user to navigate through related charts, data, and documentation will give the user the option of continuing with other tasks instead of having to just sit and wait.

B. Related Work

Computed Atlas of Surface Topography of proteins (CASTp)[16] is an online tool that locates and measures pockets on 3D protein structures. The CASTpuses a program named CAST [17]to locate and quantify pockets. The CAST employs computational geometry of complex shapes, based on alpha shape and discrete flow theory, for pocket calculation. More recently, the CASTp[18] provides annotations derived from the Protein Data Bank (PDB), Swiss-Prot, and Online Mendelian Inheritance in Man (OMIM).

The PPV and the CASTp are similar in that both display proteins and their respective pockets in Jmol, and both display protein and pocket metadata information. However, there are significant differences between the PPV and the CASTp. First, the PPV uses a different pocket extraction algorithm from the theCASTp. In many cases, the CASTp requires the user to upload PDB structured files in order to generate and visualize pockets while PPV provides on-the-fly generation and visualization of pockets given a valid PDB Id or a well-formed PDB file. The PPV also employs caching management to avoid expensive computation for pocket extraction when possible.The PPV uses AJAX for long running tasks, such as pocket generation and PDB metadata retrieval, relieving the user from needlessly waiting for page refresh.

III. Design and Implementation

The PPV project employs the Model 2 design pattern to ensure the responsibility of a particular component does not overreach into other components. We first describethe high level architecture and data flows of the PPV.Subsequent subsection takes a closer inspection of the view, controller, and model components.

A. PPV Architecture

The PPV consists of a client web interface for user inputs and visualization, a middle-layer for controlling the flow, and the backend web services performing the actual computationsuch as retrieving PDB files and pocket extraction,as shown in **Error! Reference source not found.** The web interface is implemented primarily with Java Server Pages and a cascading style sheet; the middle-layer control is implemented using JavaScript and Java servlets running on the Tomcat web server[19]; the backend web services and business logic are implemented with Java and C++ classes.

In the case of the PPV,the business logic is the preparation, extraction, and retrieval of protein PDB files and their associated pockets into PMESH [20] file format, and metadata describing those files. The PPV architecture uses a web serviceto enable the Java web client to initiate protein pocket generation implemented in C++. The reasons for choosing this architecture for the protein pocket viewer are threefold. First, using Java technologies for the web client allows the developer to capitalize on the capabilities that are robust and freely available. Second, existing program to extract the protein pockets was implemented in C++. A web service allows us to bridge the capability between the Java client and the C++ server in a clean, standards based manner. Third, packaging the pocket extraction as a web service also abstracts it as a single purpose, loosely coupled component that can be tested separately and used by others. The WSDL provides all the information necessary to create a web service client to consume the pocket extraction web service, without knowing the implementation details of the pocket extraction algorithm.

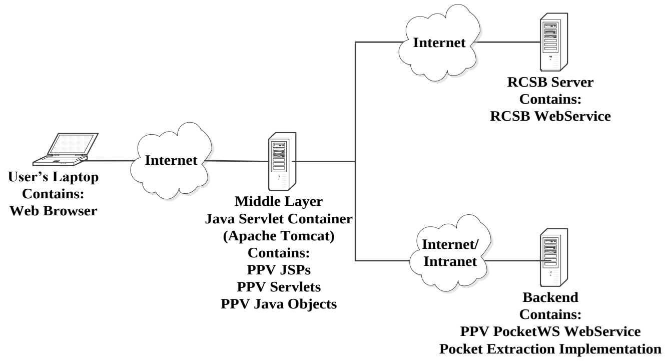


Fig. 3. Protein Pocket Viewer Web Service Architecture Diagram

The user's web browser connects to the PPV web client in the middle-layer. The PPV web client resides within an Apache Tomcat web server, containing the code for PPV JSPs, Servlets and Java objects. The PPV web client consumes the web services for protein pocket extraction, named "PocketWS", implemented in C++ and hosted on the backend computational server together with the existing pocket extraction program. The "PocketWS" web service generates the protein pocketsin the PMESH file format for displaying in the user's browser. Besides the "PocketWS" web service,the PPV web client can also consume other web services, e.g. the RCSB PdbWS(www.rcsb.org/pdb/services/pdbws) hosted on the remote RCSB server. The PdbWS web service is used by the PPV to validate PDB ID entry supplied by the user.

B. PPV DataFlow

Fig 4 shows the steps and dataflow how the PPV performs the function of extracting protein pockets and displaying them for the user.

Step 1: The user enters the PDB ID, selects the number of pockets to generate and clicks the submit button on the PPV interface, which calls the CheckPDBId servlet.

Step 2: The CheckPDBId servlet checks for the validity of the PDB ID by calling the RCSB pdbWS web service.

Step 3:TheCheckPDBId returns the validation result to the PPV interface, which displays a visual indicator if the given ID is invalid.

Step 4a:If the ID is valid, the GetPDB servlet iscalled asynchronously via AJAX.

Step 4b: The GetPocket servlet is called asynchronously via AJAX with the validated PDB Id and the number of pockets to be generated. Note that step 4a and 4b run currently in two separate threads.

Step 5: The GetPDB servlet retrieves the PDB XML file and its metadata from the RCSB PDB database, and returns it to be displayed in the PPV interface.

Step 6:The GetPocket servlet calls the "PocketWS"web service, which extracts the pocketsas PMESH files for the given protein. The PocketWS returns an array of PocketData, each of which contains pocket metadata such as volume and surface area, and the location of corresponding PMESH file. The generated pocket PMESH files are then transferred to the client and displayed in the Jmolviewer.

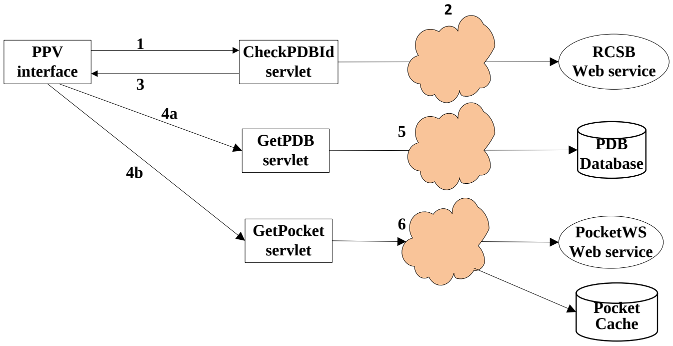


Fig. 4. The dataflow for pocket extraction and display.

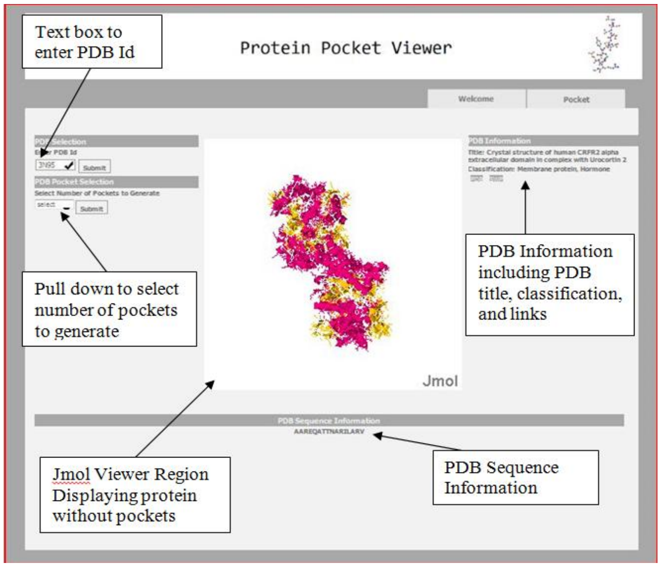
C. Implementaion

The PPV was implemented using the Model 2 design pattern. The Model 2 design consists of model, view, and controller components.

1) View Related Components:

The PPV web interface was implemented using Java Server Pages and cascading style sheets, which define the layout and look-and-feel of the main interface.

Fig 5 shows the major components in the PPV web interface. A user can entera PDB Id and select the number of pockets in order to extract and visualize the protein pockets. If the number of pockets is not selected, then no pockets will be generated, but only the PDB file will be retrieved and rendered. The PDB information region displays metadata including PDB title, classification, andhyperlink to additional information on the PDB websites. Alsothe PDB sequence region displays the amino acid sequence of the protein.

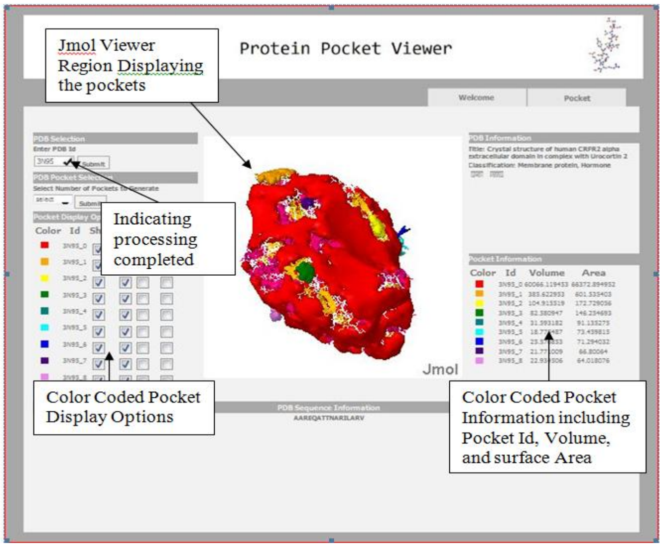

Text box to enter PDB Id
Protein Pocket Viewer
Welcome
Pocket
PDB Selection
Enter PDB Id
3N95
Submit
PDB Pocket Selection
Select Number of Pockets to Generate
select
Submit
Pull down to select number of pockets to generate
PDB Information
PDB Information including PDB title, classification, and links
Jmol
PDB Sequence Information
Jmol Viewer Region Displaying protein without pockets
PDB Sequence Information

Title: Crystal structure of human CRFR2 alpha extracellular domain in complex with Urocortin 2
Classification: Membrane protein, Hormone

AAREQATTNARILARV

Fig. 5. Major components of the PPV web interface.

Fig 6 displays the protein pockets in the Jmol viewer and additional informationassociated with the pockets. Thecolor-coded pocket display options allow the user chooses to show or hide a pocket, or change its display as a filled surface, a dot surface or a mesh. The color-coded pocket information region displays metadata (volume and surface area) of the pockets. If the protein has fewer pockets than the number selected by the user, the maximum number of available pockets will be generated and displayed.


Jmol Viewer Region Displaying the pockets
Protein Pocket Viewer
Welcome
Pocket
PDB Selection
Enter PDB Id
3N95
Submit
PDB Pocket Selection
Select Number of Pockets to Generate
select
Submit
Pocket Display Options
Indicating processing completed
Jmol
PDB Information
Pocket Information
PDB Sequence Information
Color Coded Pocket Display Options
Color Coded Pocket Information including Pocket Id, Volume, and surface Area

Title: Crystal structure of human CRFR2 alpha extracellular domain in complex with Urocortin 2
Classification: Membrane protein, Hormone

Color	Id	Volume	Area
🟥	3N95_0	60066.119433	66372.894952
🟧	3N95_1	383.622933	601.535403
🟨	3N95_2	104.915319	172.729056
🟩	3N95_3	82.380947	146.234693
■	3N95_4	31.393182	91.135275
■	3N95_5	[illegible]	73.439815
🟦	3N95_6	[illegible]	71.294032
🟪	3N95_7	[illegible]	66.80064
■	3N95_8	[illegible]	64.018076

AAREQATTNARILARV

Fig. 6. Displaying multiple protein pockets in the PPV window.

Extracting pockets for a large protein can be CPU-intensive. Since the call to the server is handled asynchronously, the user can interact with currently display while the request is being processed. An in-progress icon is placed within the PDB ID input box as a visual cue, indicatinga request being processed. Atthe successful completion of the request, a checkmark icon replaces the in-progress icon. If there is an error in the processing, e.g. an invalid PDB ID, a red 'X' icon would indicate the failure of the request. Such a visual indicator is very useful because of the asynchronous nature of request processing.

2) Controller Related Components:

The controller related components were implemented using JavaScript and Java servlet files.

There are four JavaScript files used in the current implementation. The JavaScript functions use AJAX for asynchronous execution. This asynchronous execution allows the user to interact with other components in the web application while the asynchronous calls execute. The JavaScript functions asynchronously call the appropriate servlets for services such as validating PDB ID, retrieving proteins, and extracting pockets and their metadata.

The Java servlets help to provide the controller aspect of the model-view-controller design pattern. The current PPV implementation has three Java servlet files, as shown in the middle layer of Fig 4 The servlets call appropriate methods provided by the business logic classes.

3) Model Related Components:

The PPV implemented a web service "PocketWS" for pocket extraction. The interface of the web service was defined in the "PocketWS" WSDL file, and implemented as a C++ web service class. The C++ implementation is a thin web service layer that calls upon the existing executable described in [1]with the appropriate arguments. The executable may be updated or replaced, without affecting the rest components of the PPV. Since extracting pocket is CPU intensive, the web service maintains a disk-resident cache of previously generated pockets. Atthe completion of the execution, it returns the result in XML format.

Besides the pocket extraction web service, the PPV also makes use of external web services, e.g. the RCSB web service "pdbWS" for PDB ID validation, and downloads PDB XML files using HTTP from PDB database. These business logics in the model related components wereimplementeda number of Java classes.

IV. Conclusion

The protein pocket viewer provides a web-based interface for protein pocket extraction and visualization based on the algorithm in [1]. It was implemented using web services and followed the Model 2 design pattern, consisting of a client web interface for user inputs and visualization, a middle-layer for controlling the flow, and the backend web services performing the actual CPU-intensive computation. Packaging the pocket extraction as a web service makes it as a single purpose, loosely coupled component that can be updated or replaced for a different algorithm easily. The PPV differs from other systems with features such as on-the-fly generation and visualization of pockets, asynchronous transactions via AJAX, and caching.

We found that it is effective to use web services to make existing programs available to users over the Internet with use of a modern browser.Because Java provides feature rich and widely supported graphical user interface, it was chosen to implement the web interface of the PPV.The existing executable of pocket extraction was written in C++. The web service solution bridged the cross domain and cross language solution of the existing C++ pocket executable with the Java web interface.

Future directions for the PPV may include the integration of related protein analysis capabilities, like a protein structure simulation, within the same architecture. A web application that provides multiple utilities sharing a common interface would make it more convenient for the user and ensure the compatibility of the analysis.

References

[1] X. Zhang and C. Bajaj, "Extraction, quantification and visualization of protein pockets.," Computational systems bioinformatics / Life Sciences Society. Computational Systems Bioinformatics Conference, vol. 6, pp. 275–86, Jan. 2007.

[2] J. D. Berman, "Structural properties of acetylcholinesterase from eel electric tissue and bovine erythrocyte membranes.," Biochemistry, vol. 12, no. 9, pp. 1710–5, May 1973.

[3] N. Unwin, "Refined structure of the nicotinic acetylcholine receptor at 4A resolution.," Journal of molecular biology, vol. 346, no. 4, pp. 967–89, Mar. 2005.

[4] W3C Working Group, "Web Services Architecture." [Online]. Available: http://www.w3.org/TR/ws-arch/.

[5] "Jmol: an open-source Java viewer for chemical structures in 3D." [Online]. Available: http://jmol.sourceforge.net/.

[6] H. M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T. N. Bhat, H. Weissig, I. N. Shindyalov, and P. E. Bourne, "The Protein Data Bank.," Nucleic acids research, vol. 28, no. 1, pp. 235–42, Jan. 2000.

[7] J. J. Garrett, "Ajax: A New Approach to Web Applications - Adaptive Path." [Online]. Available: http://www.adaptivepath.com/ideas/ajax-new-approach-web-applications.

[8] W3C, "Extensible Markup Language (XML) 1.0 (Fifth Edition)." [Online]. Available: http://www.w3.org/TR/REC-xml/.

[9] W3C, "SOAP Version 1.2 Part 1: Messaging Framework (Second Edition)." [Online]. Available: http://www.w3.org/TR/soap12-part1/#intro.

[10] W3C, "Web Services Description Language (WSDL) Version 2.0 Part 0: Primer." [Online]. Available: http://www.w3.org/TR/wsdl20-primer/.

[11] B. Basham, K. Sierra, and B. Bates, Head First Servlets and JSP. O'Reilly Media, 2004, p. 888.

[12] J. Zeldman, Designing With Web Standards. New Riders, 2003, p. 436.

[13] R. Harris, Murach's JavaScript and DOM Scripting. Mike Murach & Associates, Incorporated, 2009, p. 764.

[14] R. York, Beginning JavaScript and CSS Development with JQuery. Wiley, 2009, p. 529.

[15] G. Seshadri, "Understanding JavaServer Pages Model 2 architecture - Exploring the MVC design pattern," JavaWorld.com, 1999. [Online]. Available: http://www.javaworld.com/javaworld/jw-12-1999/jw-12-ssj-jspmvc.html.

[16] T. A. . Binkowski, "Castp: computed atlas of surface topography of proteins," Nucleic Acids Research, vol. 31, no. 13, pp. 3352–3355, Jul. 2003.

[17] J. Liang, H. Edelsbrunner, and C. Woodward, "Anatomy of protein pockets and cavities: measurement of binding site geometry and implications for ligand design.," Protein science : a publication of the Protein Society, vol. 7, no. 9, pp. 1884–97, Sep. 1998.

[18] J. Dundas, Z. Ouyang, J. Tseng, A. Binkowski, Y. Turpaz, and J. Liang, "CASTp: computed atlas of surface topography of proteins with structural and topographical mapping of functionally annotated residues.," Nucleic acids research, vol. 34, no. Web Server issue, pp. W116–8, Jul. 2006.

[19] "Apache Tomcat." [Online]. Available: http://tomcat.apache.org/.

[20] "Jmol Documentation." [Online]. Available: http://jmol.sourceforge.net/docs/JmolUserGuide/.

Generating an Educational Domain Checklist through an Adaptive Framework for Evaluating Educational Systems

Roobaea S. AlRoobaea
Faculty of Computing and Information Systems, Taif University, Saudi Arabia, & School of Computing Sciences,
University of East Anglia, UK

Ali H. Al-Badi
Department of Information Systems,
Sultan Qaboos University, Oman

Pam J. Mayhew
School of Computing Sciences
University of East Anglia, UK

***Abstract*—The growth of the Internet and related technologies has enabled the development of a new breed of dynamic websites that is growing rapidly in use and that has had a huge impact on many businesses. One type of websites that have been widely spread and are being widely adopted is the educational websites. There are many forms of educational websites, such as free online websites and Web-based server software. This creates challenges regarding their continuing evaluation and monitoring in order to measure their efficiency and effectiveness, to assess user satisfaction and, ultimately, to improve their quality.**

The lack of an adaptive usability checklist for improvement of the usability assessment process for educational systems represents a missing piece in 'usability testing'. This paper presents an adaptive Domain-Specific Inspection (DSI) checklist as a tool for evaluating the usability of educational systems. The results show that the adaptive educational usability checklist helped evaluators to facilitate the evaluation process. It also provides an opportunity for website owners to choose the usability area(s) that they think need to be evaluated. Moreover, this method was more efficient and effective than user testing (UT) and heuristics evaluation (HE) methods.

Keywords—Heuristic evaluation (HE);User Testing (UT);Domain Specific Inspection (DSI); Adaptive Framework; Adaptive Checklist

I. INTRODUCTION

It is clear that Heuristic Evaluation (HE) and User Testing (UT) are the most important traditional usability evaluation methods for ensuring system quality and usability (Lindgaard and Chattratichart, 2007). Currently, complex computer systems, mobile devices and their applications have made usability evaluation methods more critical; however, usability differs from one product to another depending on product characteristics. It is clear that users have become the most important factor impacting on the success of a product; if a product is produced and is then deemed not useful by the end-users, it is a failed product; nobody can use it and the company cannot make money (Nielsen, 2001). Nayebi et al., (2012) asserted, "Companies are endeavoring to understand both user and product, by investigating the interactions between them".

Traditional usability measures of effectiveness, efficiency and satisfaction are not adequate for the new contexts of use (Zaharias and Poylymenakou, 2009). HE has been claimed to be too general and too vague for evaluating new products and domains with different goals; It can produce a large number of false positives, and it is unlikely to encompass all the usability attributes of user experience and design in modern interactive systems (Chattratichart and Lindgaard, 2008). UT has been claimed to be costly, time consuming, prone to missing consistency problems and subject toenvironmental factors (Oztekinet et al., 2010). To address these challenges, many frameworks and models have been published to update usability evaluation methods (UEMs) (Alias et al., 2013); however, these frameworks and models are not applicable to all domains because they were developed to deal with certain aspects of usability in certain areas (Coursaris and Kim, 2011).

The adaptive framework was originally constructed and then evaluated in both the educational domain and social networks domain to generate domain specific-context inspection (DSI) method; in those experiences, it delivered interesting results by discovering more real usability problems in specific usability areas than HE or UT (AlRoobaea et al., 2013a) (AlRoobaea et al., 2013b). An adaptive checklist based upon the DSI method for facilitating the educational evaluation process was developed. The main objective of this paper is to address the challenges that were raised and to present this checklist which can be applied to any system in the educational domain as a tool that can be used by designers, developers, instructors, and website owners to design an interactive interface or assess the quality of existing systems. It also allows anyone to adopt any area of usability or any principle to determine the usability problems related to the five specific areas in educational system.

This paper is organized in the following way. Section 2 starts with a brief literature review including a summary of the adaptive framework. Section 3 presents the adaptive DSI checklist. Section 4presents a discussion of the findings. Section 5 presents the conclusion and future work.

II. LITERATURE REVIEW

A. Background and Motivation

The primary concern of interaction design is to develop interactive products or technologies that are usable. This means that the products should be easy to learn, effective to use, and offering a pleasurable user experience. Basically, a website is a product, and the quality of a product takes a significant amount of time and effort to develop. Web design is a key factor in determining the success of any website, and users should be the priority in the designers' eyes because usability problems in a website can have serious ramifications, over and above failing to meet the users' needs (Chen and Macredie, 2005). A high-quality product is one that provides all the main functions in a clear format, and that offers good accessibility and a simple layout to avoid users spending more time learning how to use it than satisfying their needing; these are the fundamentals of the 'usability' of a product. Poor product usability may have a negative impact on various aspects of the organization, and may not allow users to achieve their goals efficiently, effectively and with a sufficient degree of satisfaction [ISO, 1998]. The website consultants and marketing sectors have understood the number of hits, customer return rate, and customer satisfaction are extremely affected by the usability of a website [Rogers et al., 2007].

Designing interactive products and evaluating them are common stages of product development. However, the current traditional usability methods to measure effectiveness, efficiency and satisfaction are not adequate for the new contexts of use, and are not stable in the modern dynamic environment (Mankoff et al., 2003); Several studies have emphasized the importance of developing new kinds of usability evaluation methods and of constantly improving and making modifications to existing methods as a matter of priority, in order to increase their effectiveness (Guo et al., 2011). Having extensively reviewed the existing literature on web usability evaluation methods; this research is unique in systematically constructing an adaptive framework that is applicable across numerous domains. This DSI framework generates DSI checklist / tool for assessing and improving the usability of a product.

B. Description of the Adaptive Framework

The adaptive framework was developed according to an established methodology in HCI research (AlRoobaea et al., 2013a); (AlRoobaea et al., 2013b). It consists of four development steps as follows:

Development Step One (D1: Familiarization): This stage starts by justifying the need to develop a method that is specific, productive, useful, usable, reliable and valid, which can be used to evaluate an interface design in the chosen domain. It entails reviewing all the published material in the area of UEMs but with a specific focus on knowledge of the chosen domain. Also, it seeks to identify an approach that would support developers and designers in thinking about their design from the intended end-users' perspective.

Development Step Two (D2: User Input): This stage consists of mini-user testing (task scenarios, think aloud protocol and questionnaire). Users are asked to perform a set of tasks on a typical domain website and then asked to fill out a questionnaire. The broad aim of this stage is to elicit feedback on a typical system from real users in order to appreciate the user perspective, to identify requirements and expectations and to learn from their errors. Understanding user needs has long been a key part of user design, and so this step directly benefits from including the advantages of user testing.

Development Step Three (D3: Expert Input): This stage aims to consider what resources are available for addressing the need. These resources, such as issues arising from the mini-user testing results and the literature review, require a discussion amongst experts (in the domain and/or usability) in order to obtain a broader understanding of the specifics of the prospective domain. Also, it entails garnering more information through conversations with expert evaluators to identify the areas/classification schemes of the usability problems related to the selected domain from the overall results. These areas provide designers and developers with insight into how interfaces can be designed to be effective, efficient and satisfying; they also support more uniform problem description and they can guide expert evaluators in finding real usability problems, thereby facilitating the evaluation process by judging each area and page in the target system.

Development Step Four (D4: Draw Up DSI: data analysis): The aim of this step is to analyse all the data gathered from the previous three. Then, the DSI method will be established (as guidelines or principles) in order to address each area of the selected domain.

III. RESEARCH METHODOLOGY

A. Evaluation of the Adaptive Framework to Generate the DSI Adaptive Checklist

In the first stage, the researchers conducted a literature review on the materials relating to usability and UEMs as well as on the requirements of educational websites. In stage two, a mini-user testing session was conducted through a brief questionnaire that consists of four tasks, which were sent to ten users who are regular educational website users. In stage three, a focus group discussion session was carried out with eight experts in usability and the educational domain (i.e. single and double experts). Cohen's kappa coefficient was used on the same group twice to enable a calculation of the reliability quotient for identifying usability problem areas. In stage four, the researchers analysed the results of the previous three stages and incorporated findings. The intra-observer test-retest using Cohen's kappa yielded a reliability value of 0.8, representing satisfactory agreement between the two rounds. After that, the usability problem areas were identified to facilitate the process of evaluation and analysis, and to help designers and programmers to identify the areas in their systems that need improvement. Then, the DSI method was established, closely focused on educational websites, taking into an account what is called "learner-centred design". The DSI method was classified according to the usability problem areas, and checklist was developed, as shown in Table 1 in the appendix.

B. Piloting the Adaptive Checklist

A pilot study was conducted by two independent evaluators. They checked the adaptive checklist by applying it in a real experiment to make sure that there were no spelling or grammatical errors and no ambiguous words or phrases, and that all of the sentences in the adaptive checklist were sufficiently clear to be used by the evaluators. A fewer minor improvements were made,

C. Selection of the Targeted Websites

The researchers selected free educational websites (Bhargava et al., 2013). The selection process of the websites was criteria-based; 6 aspects were determined and verified for each website to achieve the research aim, and these are: 1) Good interface design,2) Rich functionality, 3) Good representatives of the free educational websites, 4) Not familiar to the users, 5) No changes will occur before and during the actual evaluation, and 6) Completely free educational websites. In order to achieve a high level of quality in this research, the researchers chose three well-known websites in this domain that each has all the aspects mentioned above (skoool, AcademicEarth, BBC KS3bitesize).

D. Actual Evaluation

After constructing the DSI checklist, the researchers test it intensively through rigorous validation methods to verify the extent to which it achieves the identified goals, needs and requirements that the adaptive DSI checklist was originally developed to address. It was conducted alongside heuristics evaluation (HE) and user testing (UT). The aim of this process is to collect data ready for analysis (analytically, empirically, and statistically).

Therefore, 8 expert evaluators were recruited to use the adaptive DSI checklist and HE checklist that had been developed by the researchers to facilitate the evaluation process for both methods. The evaluators had selected from the adaptive DSI checklist the usability areas and the appropriate principles for each website. Then, the actual expert evaluation was conducted and the evaluators evaluated all websites consecutively, rating all the problems they found in a limited time (which was 90 minutes). After that, they were asked to submit their evaluation report, and to give feedback on their own evaluation results. Next, 60 users were recruited for using UT. Before starting the actual evaluation, all users were given a UEM training pack. Each user was given the task scenario sheet and asked to read and then perform one task at a time.

IV. Discussion and Findings

The researchers extracted the problems discovered by the three methods from the problems sheet and removed all false positive problems, subjective problems, and duplicated problems during the debriefing session. The problems agreed upon were merged into a unique master problem list (see Table 1, Table 2, and Table 3), and any problems upon which the evaluators disagreed were removed.

TABLE I. Total Problems Found In Bbc Ks3bitesize

Method Problem type	UT	HE	DSI	Total problems
Catastrophic	0 (0%)	0 (0%)	1 (100%)	1
Major	2 (66%)	0 (0%)	3 (100%)	3
Minor	5 (100%)	0 (0%)	5 (100%)	5
Cosmetic	9 (100%)	2 (22%)	2 (22%)	11
No. problems	16 (80%)	2 (10%)	12(60%)	20

TABLE II. Total problems found in Skoool

Method Problem type	UT	HE	DSI	Total problems
Catastrophic	1 (25%)	2 (50%)	4 (100%)	4
Major	3 (30%)	2 (20%)	6 (89%)	7
Minor	2 (29%)	3 (43%)	11 (85%)	11
Cosmetic	7 (54%)	3 (23%)	12 (92%)	12
No. of problems	13 (38%)	10 (29%)	33 (97%)	34

TABLE III. Total problems found in Academic Earth

Method Problem type	UT	HE	DSI	Total problems
Catastrophic	0 (0%)	1 (33%)	3 (100%)	3
Major	3 (50%)	3 (50 %)	4 (66%)	6
Minor	2 (17%)	7 (58 %)	11 (92%)	12
Cosmetic	7 (50%)	2 (14%)	11 (79%)	14
No. of problems	12 (34%)	13 (37%)	29 (83%)	35

Generally, UT, HE and adaptive DSI checklist revealed different types and numbers of usability problems. One-way ANOVA reveals that there is significant difference between the three methods in terms of discovering usability problems on the whole ($F = 13.447$, $p < 0.001$). UT, HE and the adaptive DSI checklist revealed 80%, 10% and 60% of the real usability problems found in the BBC KS3bitesize website, respectively. One-way ANOVA-Tukey HSD was used and the results show that there is a strongly significant difference amongst the methods in finding usability problems on the BBC KS3bitesize website between HE and UT, where $p = 0.003$. In the Skoool website, UT, HE and the adaptive DSI checklist revealed 38%, 29% and 97% of the found real usability problems, respectively. One-way ANOVA-Tukey HSD was used and the results show that there is a strongly significant difference amongst the methods in finding usability problems in Skoool (as a dependent factor), particular between HE and the adaptive DSI checklist and between the adaptive DSI checklist and UT, where $p < 0.001$. Finally, UT, HE and the adaptive DSI checklist revealed 34%, 37% and 83% of the found real usability problems in Academic Earth, respectively. One-way ANOVA-Tukey HSD was used and the results show that there is a significant difference amongst the methods in finding usability problems in Academic Earth between HE and UT, where $p = 0.044$. The performance of HE in discovering real usability problems totally ranged from 10% to 37%. UT discovered real usability problems ranging from 34% to 80%, while the adaptive DSI checklist discovered real usability problems ranging from 60% to 97%. Also, UT and HE performed better in discovering major, minor and cosmetic real usability problems, but the adaptive DSI checklist was the best in discovering more catastrophic, major, minor and cosmetic real usability problems. Furthermore, 9 unique problems were discovered in all experiments on the three websites through UT (6 in BBC KS3bitesize and 3 in Academic Earth); whereas the remaining UT problems were discovered by the adaptive DSI checklist (although one was discovered by HE). Thus, it can be seen that the adaptive DSI checklist was the best in discovering real problems; UT came second, and HE in third place.

V. Conclusion And Future Work

The main aim of this experiment was to evaluate the adaptive DSI checklist for the educational websites through its ability to discover usability problems by comparing its results with usability testing (UT) and heuristic evaluation (HE). The adaptive DSI checklist seemed to guide the evaluators' thoughts in judging the usability of the websites. This finding facilitates decision-making with regard to which of these methods to employ. Also, it addresses the shortcomings of these methods; hence, to avoid wasting money and time an alternative method that is well-developed, context-specific and adaptive to the situation in hand, such as what has been generated here for the educational domain, should be employed. This research contributes to the advancement of knowledge in the HCI field by introducing the adaptive DSI checklist that is specific for evaluating educational systems. In order to consolidate and confirm the findings, future research could include testing the adaptive DSI checklist by applying it, for example, to web-based server applications. Also, we need to further test the adaptive framework by developing an adaptive DSI checklist for different fields, such as e-commerce or news sites.

REFERENCES

[1] Alias, N. Siraj, S. DeWittD. Attaran, M.and Nordin, A.(2013). Evaluation on the usability of physics module in a secondary school in Malaysia: Students retrospective",The Malaysian Online Journal of Educational Technology, p. 44.

[2] AlRoobaea, R. Al-Badi, A. Mayhew P. (2013a). Generating a Domain Specific Inspection Evaluation Method through an Adaptive Framework: A Comparative Study on Educational Websites. International Journal of Human Computer Interaction (IJHCI), Volume 4, NO. 2.

[3] AlRoobaea, R. Al-Badi, A. Mayhew P. (2013b). Generating a Domain Specific Inspection Evaluation Method through an Adaptive Framework: A Comparative Study on Social Networks. International Journal of Advanced Computer Science and Applications. Volume 4, No. 6.

[4] Bhargava, P., Dhand, S., Lackey, A. E., Pandey, T., Moshiri, M., &Jambhekar, K. (2013). Radiology Education 2.0—On the Cusp of Change: Part 2. eBooks; File Sharing and Synchronization Tools; Websites/Teaching Files; Reference Management Tools and Note Taking Applications. Academic Radiology, 20(3), 373-381.

[5] Chattratichart, J.and LindgaardG.(2008).A comparative evaluation of heuristic-based usability inspection methods. In CHI, vol. 8, pp. 05-10.

[6] Chen S.and Macredie, R. (2005).The assessment of usability of electronic shopping: A heuristic evaluation. International Journal of Information Management, vol. 25, no. 6, pp. 516-532.

[7] Coursaris,C.and Kim, D. (2011).A meta-analytical review of empirical mobile usability studies. Journal of Usability Studies, vol. 6, no. 3, pp. 117-171.

[8] Guo, Y. Proctor, R.and Salvendy, G. (2011).A conceptual model of the axiomatic usability evaluation method",Human Interface and the Management of Information. Interacting with Information, pp. 93-102.

[9] ISO (1998). ISO 9241-11: Ergonomic Requirements for Office Work with Visual Display Terminals (VDTs): Part 11: Guidance on Usability.

[10] Lindgaard G., and Chattratichart,J. (2007). Usability testing: what have we overlooked?" in Proceedings of the SIGCHI conference on Human Factors in Computing Systems, pp. 1415-1424, ACM.

[11] Mankoff, J. Dey, A. Hsieh, G. Kientz, J. Lederer S.and Ames, M.(2003).Heuristic evaluation of ambient displays", in Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, pp. 169-176, ACM.

[12] Nayebi, F. DesharnaisJ., and Abran, A.(2012).The state of the art of mobile application usability evaluation", in Electrical & Computer Engineering (CCECE), 2012 25th IEEE Canadian Conference on, pp. 1-4, IEEE.

[13] Nielsen, J. (2001). Did poor usability kill e-commerce,in www.useit.com.

[14] Oztekin, A. Kong Z.and Uysal, O. (2010).Uselearn: A novel checklist and usability evaluation method for e-learning systems by criticality metric analysis.International Journal of Industrial Ergonomics, vol. 40, no. 4, pp. 455-469.

[15] Rogers, Y., Sharp, H., and Preece, J. (2007). Interaction design: beyond human-computer interaction. Wiley.

[16] Zaharias P., and Poylymenakou, A. (2009). Developing a usability evaluation method for e-learning applications: Beyond functional usability,Intl. Journal of Human-Computer Interaction, vol. 25,no. 1, p. 75-98.

Table 1: The adaptive Domain Specific Inspection checklist for evaluating educational system usability

Usability problem area	The adaptive Domain Specific Inspection (DSI) checklist
User usability	*Supports modification and progress of evaluation:* ○ *Does the system make important keys larger than other keys?* ○ *Does the system anticipate the user's next activity correctly?* ○ *Does the system allow the user to initiate actions?* ○ *Does the system provide an overview of the work process that has been completed by the user?*
	Supports user tasks and avoids difficult concepts: ○ *Does the system provide constructive, brief, unambiguous descriptions of the task when needed?* ○ *Does the system match the menu structure to the task structure? Can the user distinguish between options and content on the pages? Are there breadcrumbs to show where the user is and where the user last was?* ○ *Does the system use clear, simple language for questions and answers?* ○ *Does the system provide correct spelling and grammar, and understandable graphic symbols?* ○ *Does the system provide the minimal number of clickable actions, infrequent selection, and infrequent scrolling to complete one main task? Are lesson pages easy to bookmark?* ○ *Is an item visible when it should be hidden from the view, and vice versa?*
	Feedback and support services: ○ *Is feedback given at any specific time tailored to the content or problem being studied by the user?* ○ *Does feedback provide the user with meaningful information concerning their current level of achievement within the program? Is any message of current status related to the user's task* ○ *Does the system program provide the user with opportunities to access extended feedback from instructors through email and Internet communication? Is adequate FAQ offered?* ○ *Dothe performance support tools provided mimic their real–world counterparts?*
	Error Prevention: ○ *Do error messages prevent potential errors from happening?* ○ *Does the system provide solutions that help the user recover from errors, such as providing undo and redo features?* ○ *Can errors be averted or minimized when possible?*
	Easy to remember: ○ *Is a casual user able to return to using the system after some period without having to learn everything all over again? Are all functions and information well presented to support memorability?*
Motivational factors	*Supports leaner curiosity:* ○ *Does the system support the user's cognitive curiosity through surprises, paradoxes and humour, and does it deal with topics that are already of interest to the user?*
	Learning content design and Attractive screen design: ○ *Are the vocabulary and the terminology used appropriate and presented with good background, giving suitable examples?* ○ *Is the organization of the content pieces and learning objects suitable for achieving the primary goals of the system?* ○ *Are similar learning objects organized in a similar style?* ○ *Is the screen layout efficient and visually pleasing (it should appear simple, i.e., uncluttered, readable and memorable)?* ○ *Are the font choices, colours and sizes consistent with good user screen design?*
	Motivation to learn: ○ *Does the system use e-stories, simulations, discussion messages, role-playing and activities to gain the attention and to maintain the motivation of the user to learn more?* ○ *Does the system provide the user with frequent and varied learning activities that increase learning success?* ○ *Are the user's actions rewarded by audio, video, text or animations, and are the rewards meaningful?* ○ *Is the system easy to learn but hard to master? Is the system paced to apply pressure but not frustrate the user? Does the difficulty level vary so that users are given greater challenges as they develop mastery?* ○ *Is the user's fatigue minimized by varying the activities and the difficulty levels during a learning session?*
Content information and process orientation	*Relevant, correct and adequate information:* ○ *Does the system display only information that is relevant to its purposes?* ○ *Does the system update the content constantly?* ○ *Does the system display only the available lesson,and is the content suitable to the page length?* ○ *Does the system provide concise and non-repetitive information?* ○ *Does the system offer an amount of information that isappropriate to the page length, and is all the text of a viewable/readable size?*
	Reliability and Validity: ○ *Is there a link provided to the homepage? Was the system was built by a reliable institution?* ○ *Are reliability, stability and continuity of learning in the system guaranteed?*

	Privacy and Security: ○ *Are sensitive areas protected by passwords and an SSL protocol (e.g., VeriSign™) against hackers?*
Learning process	*Assessment:* ○ *Does the system include self-assessment for each module, whether in audio, video or text, and does it keep a record of progress?* ○ *Does the system provide sufficient feedback (audio, video) to the user in order to provide corrective directions?* ○ *Does the system provide the instructor with the user's evaluation and tracking reports?*
	Interactivity: ○ *Does the user become engaged with the system program through activities that are challenging? Is the presentation of the lessons designed to promote engagement?* ○ *Is the user able to respond to the program at leisure?* ○ *Does learning become easier with an interactive approach, wherein users quickly learn how best to respond to the program? Does the user gain in confidence by so doing?* ○ *Does the user have confidence that the system is interacting and operating in the way it was designed to?*
	Evokes mental images for the users: ○ *Does the system allow the user to use their imagination, in a way that enhances their comprehension?* ○ *Does the system appeal to the imagination and does it encourage recognition in order for the user to create unique interpretations of the characters or contexts?* ○ *Is the user interested in the system characters because they are drawn from the user's own culture?*
	Resources: ○ *Does the system provide access to a wide range of resources (e.g., examples and real data archives) appropriate to the learning context?* ○ *If the system includes links to external www. Or to intranet resources, are the links kept up to date?*
	Learning management: ○ *Can the user manage all the activities pertaining to the learning program with ease? Can the user clearly understand everything, and perceive options for additional guidance (chat, edit, add, seek instruction or other forms of assistance) when needed? Are all control items logically labelled and grouped in a control panel?* ○ *Are the lessons easy to upload, download, share, retrieve and organise? Do the lessons support various learning styles, and do they support synchronous and asynchronous modes.*
	Learnability: ○ *Is the system designed such that the user finds it easy to learn how to use?*
Design and media usability	*Multimedia representations:* ○ *Does multimedia help the user in all aspects to learn interactively by playing videos, audio files and audio mock tests, and does it make learning enjoyable?* ○ *Does the system include sound and visual effects? Do these effects provide meaningful feedback or hints, and do they stir particular emotions?* ○ *Does the system include surprises, humour and interesting representations for the user, and does it avoid unnecessary multimedia representations that could confuse a user who has just started to work with the system?* ○ *Is the user allowed to skip non-playable and frequently repeated content in videos or learning games?* ○ *Is the user allowed to customize video and audio settings, and to adjust the difficulty level?*
	Accessibility and compatibility of hardware devices: ○ *Is the system compatible with various platforms and hardware? Are its features adaptable to individual user preferences?* ○ *Do potential users have all the necessary computer skills to use the application? (There should be consistency between the motor effort and skills required by the hardware and the developmental stage of the learner audience.)* ○ *Are all input devices/buttons that have no functionality disabled to prevent user-input errors?* ○ *Are the lessons accessible to users with physical impairments, and are their contents available in various languages?*
	Functionality: ○ *Is all the necessary functionality of the system available without having to leave the site, and does it work correctly?* ○ *Is all functionality clearly labelled, and does it facilitate easy task completion? Is the system status of each task clear on all pages?*
	Navigation and Visual clarity: ○ *Are navigation objects and tools kept in particular, clearly defined positions, and are they of an adequately viewable size?* ○ *Is unnecessary animation and Flash avoided?* ○ *Is content logically structured in different sections and levels with enough space between the individual items? Are the colours and graphics used suitable for promoting navigation?* ○ *Are the menus understandable and straightforward, and are the items logically grouped and labelled? Do all buttons, links and features have a 'mouseover' or pop-up window that provides meaningful feedback?* ○ *Is a site map and/or a table of contents available, as well as a calendar?* ○ *Is the site navigation consistent, and is the search engine accurate?*

	o *Is the user’s current position in the system clearly labelled, and are adequate ‘back buttons’(to previous pages) provided? Can the user clearly identify where to start on the system’s Homepage?* o *Is the need for scrolling down a page kept to a minimum?* o *Are all functions, buttons and links labelled meaningfully, and are their intended functionalities clear?*

Semantic, Automatic Image Annotation Based On Multi-Layered Active Contours and Decision Trees

Joanna Isabelle OLSZEWSKA
School of Computing and Engineering
University of Huddersfield
Queensgate, Huddersfield, HD1 3DH, UK

Abstract—**In this paper, we propose a new approach for automatic image annotation (AIA) in order to automatically and efficiently assign linguistic concepts to visual data such as digital images, based on both numeric and semantic features. The presented method first computes multi-layered active contours. The first-layer active contour corresponds to the main object or foreground, while the next-layers active contours delineate the object's subparts. Then, visual features are extracted within the regions segmented by these active contours and are mapped into semantic notions. Next, decision trees are trained based on these attributes, and the image is semantically annotated using the resulting decision rules. Experiments carried out on several standards datasets have demonstrated the reliability and the computational effectiveness of our AIA system.**

Keywords—automatic image annotation; natural language tags; decision trees; semantic attributes; visual features; active contours; segmentation; image retrieval

I. Introduction

With the increasing amount of available visual digital data, labeling [21] or searching [24] for an image remains a challenging task, not only because it necessitates a computationally efficient management of image storage and indexing processes, but also it requires the investigation of the semantic gap, i.e. the difference between the visual image representation and its linguistic description.

For this purpose, several image retrieval (IR) techniques have been developed in the literature. In the tag-based retrieval approach, images are retrieved on the basis of the textual information which has been beforehand manually associated to the images, whereas in the content-based image retrieval (CBIR) method, images are retrieved on the basis of low-level visual information automatically extracted from the images [1].

In this work, we focus on the most recent approach called Automatic Image Annotation (AIA), whose main steps are the automatic extraction of visual features from images and their automatic, semantic labeling. This latter step usually requires a training to learn the semantic concepts from image samples and to use these concepts to label new images. Thus, these images, which are automatically annotated with semantic labels, can be retrieved by users providing keywords such as in the tag-based retrieval approach rather than a query image as it is the case for CBIR. Hence, AIA combines the advantages of both tag-based and content-based image retrieval approaches.

Whereas most of the image annotation approaches are still manual [6] for both object delineation and labeling such as LabelMe [27], some automatic image annotation techniques have been recently developed [32].

AIA systems mainly use graph-based algorithms, e.g. Normalized cut (N-cut) [29] or region growing methods [5], as segmentation methods. In general, these approaches are appropriate for segmenting background objects, but not a main object itself, since they usually tend to oversegment the studied image. This results in the loss of the main object's entirety and in a mix of foreground's parts with the background ones. Hence, the features extracted from these resulting regions are not specific enough to characterize the main object. The active contour approach [10] does not present this drawback as it delineates the boundaries of the entire object. However, active contours have been used up to now only for semi-automatic graphic annotation processes [8], [17], [9], thus not providing fully automatic graphic nor semantic annotations, as these specific implementations found in the literature present weaknesses in presence of noise and do not offer any semantic computational framework.

In AIA, the semantic labeling of images usually implies the use of classifiers such as artificial neural network (ANN) [25], [11], or support vector machine (SVM) [4], [8], but these methods require computationally expensive training. Decision trees (DT) have been proven to be much faster and allow both categorical and numerical values [18]. The classification could thus rely on semantic rules and visual features.

In this paper, we propose a new fully automatic image annotation method based on efficiently implemented active contours and decision trees. Hence, our approach consists of the automatic recursive image segmentation in multiple layers using multi-feature active contours and the automatic semantic labeling of the image based on decision trees.

While being an unsupervised segmentation technique, the multi-layered multi-feature active contour approach does not

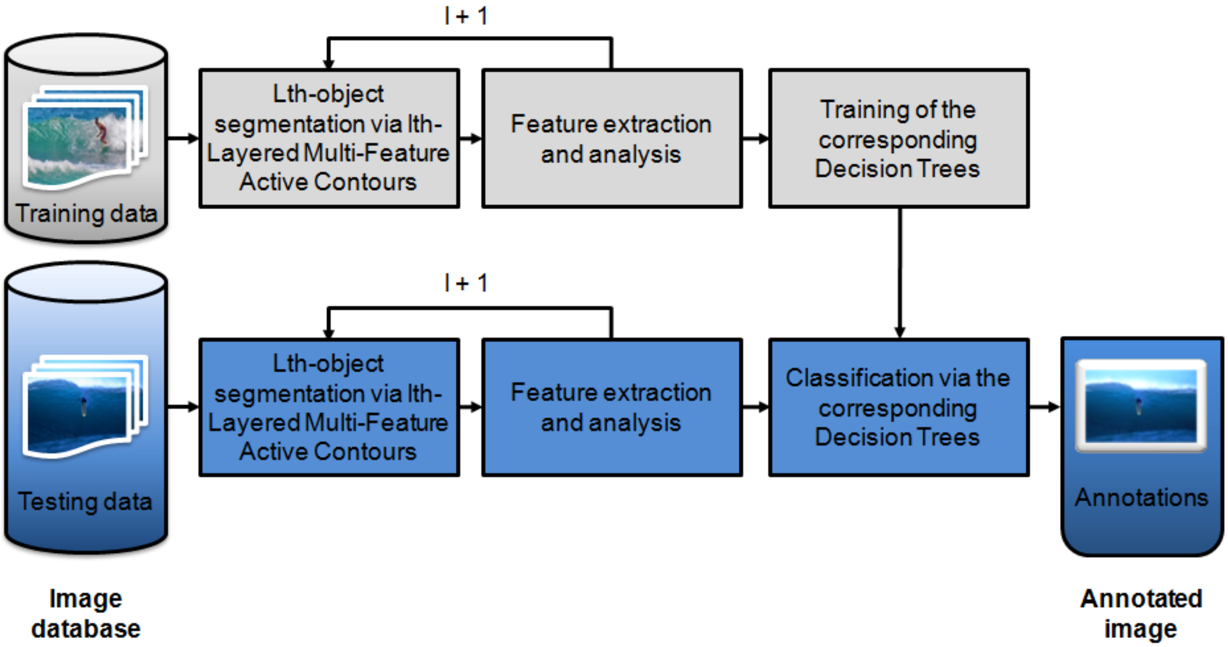

l + 1
Lth-object segmentation via lth-Layered Multi-Feature Active Contours
Feature extraction and analysis
Training of the corresponding Decision Trees
Training data
l + 1
Lth-object segmentation via lth-Layered Multi-Feature Active Contours
Feature extraction and analysis
Classification via the corresponding Decision Trees
Testing data
Annotations
Image database
Annotated image

Fig. 1. Architecture of our Automatic Image Annotation process.

use any prior knowledge about the foreground unlike top-down segmentation methods [2] and reaches a semantically coherent segmentation of the objects more accurately than the bottom-up segmentation techniques [28] and faster than the combined ones [13], [14] or [12].

On the other hand, our segmentation method also provides the background region. However, in this work, we only exploit the information about the main object and its subparts, in order to process the training of the corresponding decision trees and the automatic labeling of the dataset images in a more computational efficiently way than background-based systems like [8].

AIA approaches consider usually that the main object is in the center part of the image [11] or constitutes the largest region of the image [25]. Because of these constraining assumptions on the position or importance of the main object, these systems cannot classify nor annotate an image properly if the object appears in another part of the image. This is not the case for the adopted multi-feature active contour approach which allows the detection of any object in any part of the image [19], [22].

Moreover, when compared to [7], our multi-layered multi-feature active contour method provides not only semantically coherent objects but also a semantically meaningful sub-object decomposition without any training.

The contributions of this paper are as follows:

- the use of active contours into a computationally efficient, full AIA system;
- the introduction of multi-layered active contours based on the robust and effective multi-feature active contours, in order (i) to precisely and automatically segment the image into background and semantically meaningful foreground regions and (ii) to extract coherent and semantically meaningful sub-regions of the extracted main object;
- the proposed architecture of the novel automatic image annotation process involving decision trees relying on hierarchic, semantic attributes derived from multi-stage visual features, which ones have been extracted from the image regions segmented by the corresponding multi-layered multi-feature active contours.

The paper is structured as follows. In Section II, we present our Automatic Image Annotation (AIA) approach based on the unsupervised, semantic labeling of an image under investigation, given visual features of objects extracted from the segmented image by means of multi-layer active contours and given trained decision trees. The resulting annotation system has been successfully tested on a challenging database containing real-world images with very close semantic classes as reported and discussed in Section III. Conclusions are drawn up in Section IV.

II. Our Proposed Annotation System

In this section, we describe our AIA system illustrated in Fig. 1, which performs both the automatic visual segmentation of the image and its automatic semantic annotation. The main steps of the process are the multi-layered partition of the image in terms of background, foreground and foreground´s semantically meaningful sub-regions (Section II.A), the extraction of the corresponding metric features from these delineated regions as well as the definition of the semantic attributes based on the visual features (Section II.B), and the

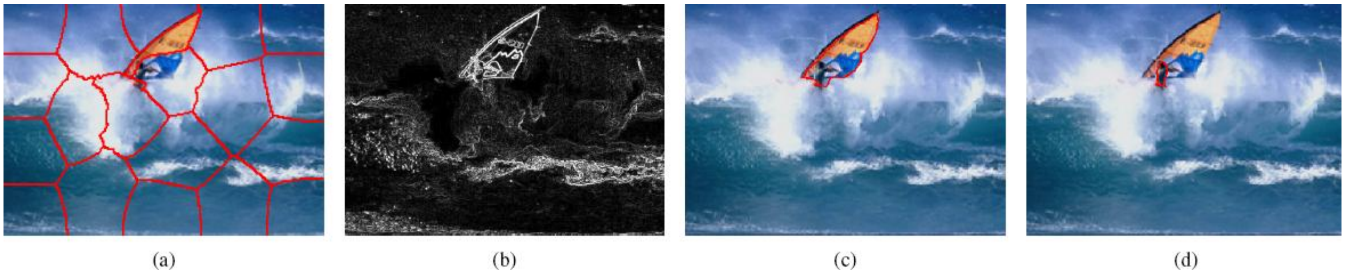


Fig. 2. Feature extraction with (a) N-cut method; (b) edge map; (c) first-layer multi-feature active contour; (d) second-layer multi-feature active contour. Best viewed in color.

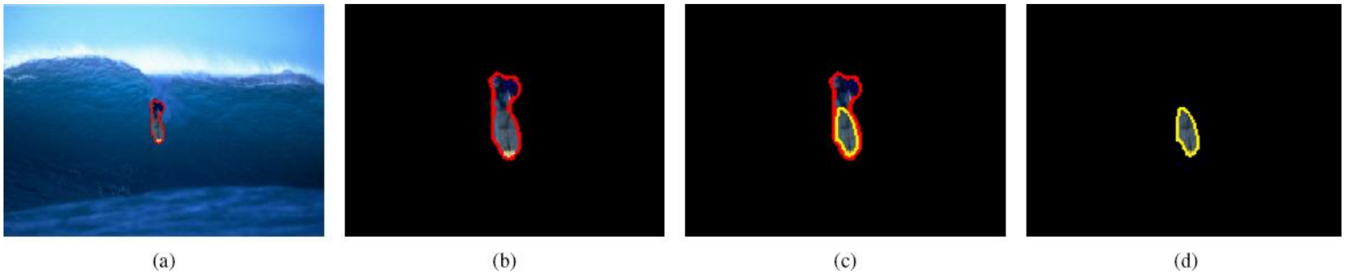


Fig. 3. Image segmentation with multi-layered multi-feature active contours: (a) main-object segmentation with the first-layer multi-feature active contour; (b) main-object extraction; (c) sub-object segmentation with the second-layer multi-feature active contour; (d) sub-object extraction. Best viewed in color.

labeling of the image followed by the final online annotation of the image using offline-trained decision trees (Section II.C).

A. *Multi-Layered Multi-Feature Active Contours*

Active contours [10] are deformable two-dimensional closed curves that evolve in the image plane from a given initial position to the foreground boundaries characterizing thus the shape and the position of the object of interest.

In this work, we have chosen multi-feature active contours [19] to segment images in order to provide visual information to the system. Multi-feature active contours are particularly suitable for image annotations, since they can precisely segment images in semantically meaningful parts. Indeed, they could extract main object(s) entirely and very efficiently as illustrated in Figs. 2(c)-(d). It is worth to note that this is not the case of most of the state-of-art segmentation methods such as N-cut [29], [3] or edge detection [31], which usually suffer from over-segmentation and do not necessarily grasp the objects of interest as shown in Figs. 2 (a)-(b), respectively.

Other major advantages of multi-feature active contours [19] are as follows:

- they use both region-based and edge-based image representation;
- they combine the positive properties of bottom-up and top-down approaches, whereas do not require any prior knowledge, in order to not constrain the contour evolution, leading to the accurate delineation of main objects with highly varying shape and appearance;
- they are robust towards noise, clutter, and complex backgrounds.

Multi-feature active contour representation consists in a parametric plane curve $\mathcal{C}(s) : [0,1] \rightarrow \mathbb{R}^2$ modeled by a B-Spline formalism, while its evolution is guided by internal forces (α: elasticity, β: rigidity) described by the curve´s mechanical properties and the external force Ξ resulting from multiple characteristics of the image under study, computed by the dynamic equation as follows:

$$\mathcal{C}_t(s,t) = \alpha\, \mathcal{C}_{ss}(s,t) - \beta\, \mathcal{C}_{ssss}(s,t) + \Xi. \tag{1}$$

The external force Ξ based on the Multi-Feature Vector Flow (MFVF) [19] has a large capture range as well as a bidirectional convergence and owns additional capacities related to the properties of the extracted features. Equation (1) sets the general framework of the multi-feature active contours, allowing the use of an extensible number of different features describing the shape and appearance of the objects of interest [19].

Multi-layered multi-feature active contours segment an image *I* into several parts or equivalently in *l* + *i* layers, namely, the background (*i* = 0), the foreground (*i* = 1) and the foreground sub-regions (*i* = 2). The segmentation is recursively performed by applying *ith*-times multi-feature active contours. In the first step, the multi-layered multi-feature active contours divide the image into background/foreground such as illustrated in Fig. 3 (a). The background corresponds to the layer *l*, while the main object or foreground F_{l+1} lies in the layer *l + 1* shown in Fig. 3 (b).

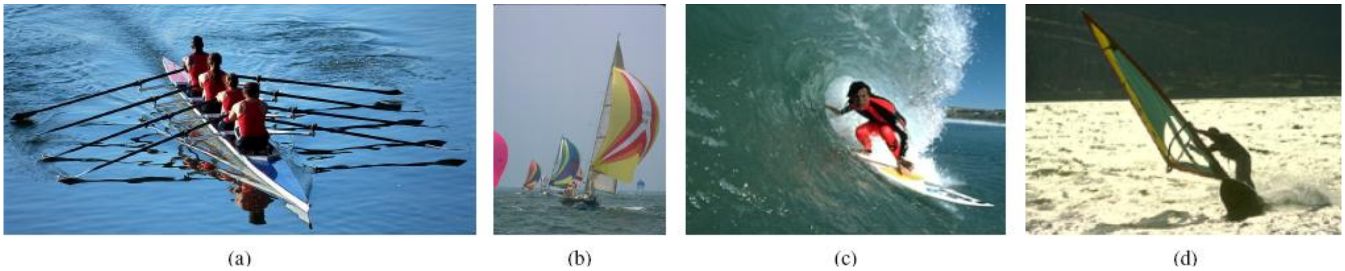


Fig. 4. Examples of four semantic classes of our dataset: (a) rowing; (b) sailing; (c) surfing; (d) windsurfing.

In the second step, the foreground is segmented again using the multi-layered multi-feature active contours as depicted in Fig. 3 (c). It results in s sub-regions or sub-parts of the main object $F_{l+2,1}$, ... $F_{l+2,s}$, with Fl+i = Fj=1,...,s Fl+i,j (i, j 2 N). Figure 3 (d) shows $F_{l+2,1}$ which is semantically meaningful. This process of image partition leads to the delineation of coherent objects, allowing efficient foreground labeling and automatic image annotation as described in the next sections.

B. Feature Extraction and Analysis

Each region $F_{l+i,j}$ segmented by the active contours at the layer *i* could be characterized by metric features such as their mean color values in the *RGB* color space. However, humans use semantic concepts to identify and describe colors [16]. Hence, we adopt both numeric features directly extracted from the image such as Region_Average_Color and semantic features like Region_Color_Name mapped from the visual features as described in [20].

Unlike [30] or [8], texture features are not considered in our work in order to allow our automatic system to annotate low-resolution and noisy images as well.

Geometric properties of the delineated regions could be described with notions such as Region_Center_of_Gravity, Region_Shape = {*oval*, *rectangular*, *triangular*}, and Region_Area. Indeed, linguistic concepts have been proven to complement well visual information in the process of scene understanding [23].

C. Decision Trees (DT)

Decision trees (DTs) [26], [15] are a form of multiple variable analysis based on multi-level decisions which split data into a hierarchy of branches that produce the characteristic inverted tree shape. Each segment or branch is called a node. Each node could be of two types, namely, internal node and terminal node also called leaf. Each internal node corresponds to a decision governed by an attribute dividing the data samples the most effectively. Each leaf represents the outcome of the data samples that follow the path from the root (top node) of the tree to the corresponding leaf. The leaves have mutually exclusive assignment rules, and thus, they can be expressed with unique $if - then$ rules, called decision rules, which are interpretable semantically.

In fact, each data sample is represented by a vector of attributes and its associated values. The discovery of the decision rules to create the branches underneath the root node is based on the extraction of the relationship between the input attributes of the samples and the outcomes. The standard DT process implies that each sample has only one possible outcome, i.e. belongs to a single class.

A DT is trained using a set of labeled samples. During the training phase, a DT is built by recursively dividing the training samples into non-overlapping sets. Every time the samples are divided, the attribute used for the division is discarded. The procedure continues until all samples of a same class reach the tree's maximum depth when no attribute remains to separate them.

The classification of new samples is done by performing a sequence of tests. Hence, during the testing phase, the DT is traversed from the root to a leaf node using the attribute values of each new sample. The decision of the sample is the outcome of the leaf node where the sample reaches.

Compared to other machine learning methods such as support vector machine (SVM) such as in [8] or Artificial Neural Networks (ANN) used in [25], DT is naturally interpretable in human language, is fast, and its learning requires only a small numbers of samples. Moreover, DT is robust for incomplete and noisy data and handles both semantic and numeric values.

In this work, we used several decision trees to achieve the goal of automatically annotating images. Our approach consists in using semantic decision rules to classify the images into classes based on their semantic attributes, which were defined using trained decision trees involving both semantic and visual features. Hence, decision trees are first induced in order to define keywords based on numeric visual features and semantic features introduced in Section II.B. Next, higher semantic level decision trees are built to classify the images into the classes based on these natural-language keywords.

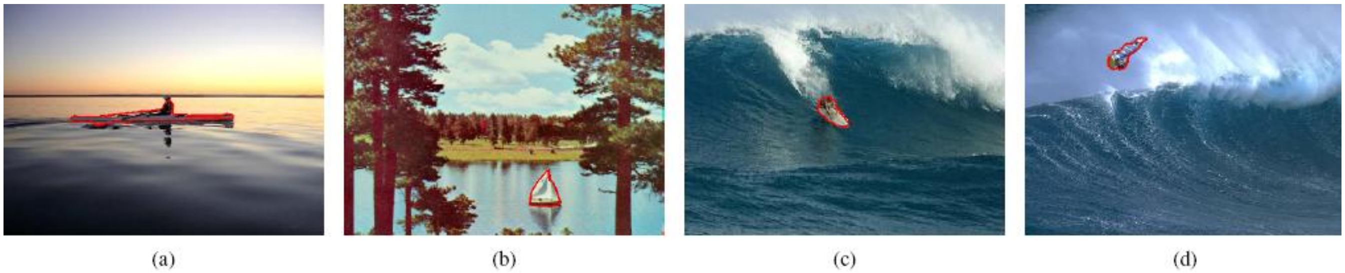


Fig. 5. Examples of automatically segmented images from the dataset. Best viewed in color.

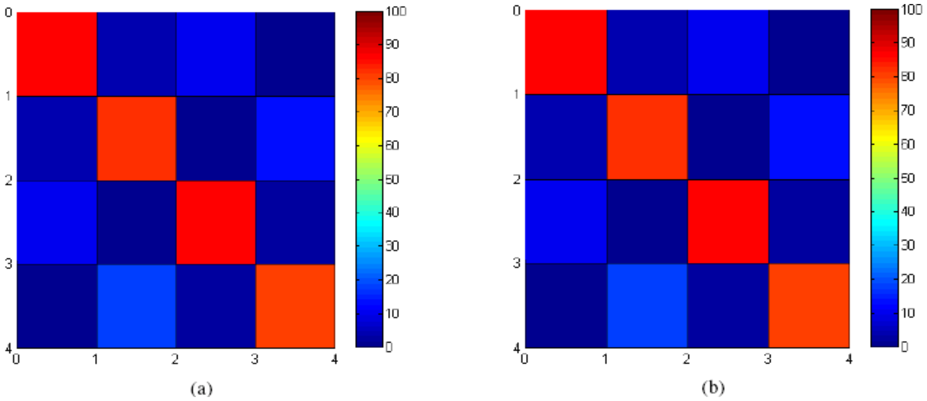


Fig. 6. Fig. 6 est viewed in color.

As an example, we use a dataset with 'water sport' images that should be automatically annotated. More information about this dataset are provided in Section III.

For this purpose, we consider at first the definition of keywords such as 'board', 'boat', 'paddle', and 'sail' based on the extracted visual features from the regions delineated by multi-layer active contours. The corresponding induced decision rules are as follows:

$$\begin{cases} if \ \text{Region\_Shape} = oval \\ \quad then \ \text{outcome} = board \\ if \ (\text{Region\_Shape} = rectangular \\ \quad and \ \text{Region\_Color\_Name} = black) \\ \quad then \ \text{outcome} = paddle \\ if \ \left(\text{Region\_Shape} = rectangular \right. \\ \quad \left. and \ \text{Region\_Area} > \dfrac{M \times N}{80}\right) \\ \quad then \ \text{outcome} = boat \\ if \ \text{Region\_Shape} = triangular \\ \quad then \ \text{outcome} = sail \end{cases} \qquad (2)$$

with M and N, the width and the height of the image under investigation, respectively.

Next, we classify the 'water sport' images of the dataset into four classes (Fig. 4), namely, 'rowing', 'sailing', 'surfing', and 'windsurfing', which are semantically closely related, by inducing a decision tree whose leaf nodes can be expressed with unique $if - then$ semantic rules as follows:

$$\begin{cases} if \ \text{paddle} = yes \\ \quad then \ \text{outcome} = rowing \\ if \ (\text{paddle} = no \ and \ \text{board} = yes \\ \quad and \ \text{sail} = yes) \\ \quad then \ \text{outcome} = windsurfing \\ if \ (\text{paddle} = no \ and \ \text{board} = yes \\ \quad and \ \text{sail} = no) \\ \quad then \ \text{outcome} = surfing \\ if \ (\text{paddle} = no \ and \ \text{board} = no \\ \quad and \ \text{boat} = yes) \\ \quad then \ \text{outcome} = sailing. \end{cases} \qquad (3)$$

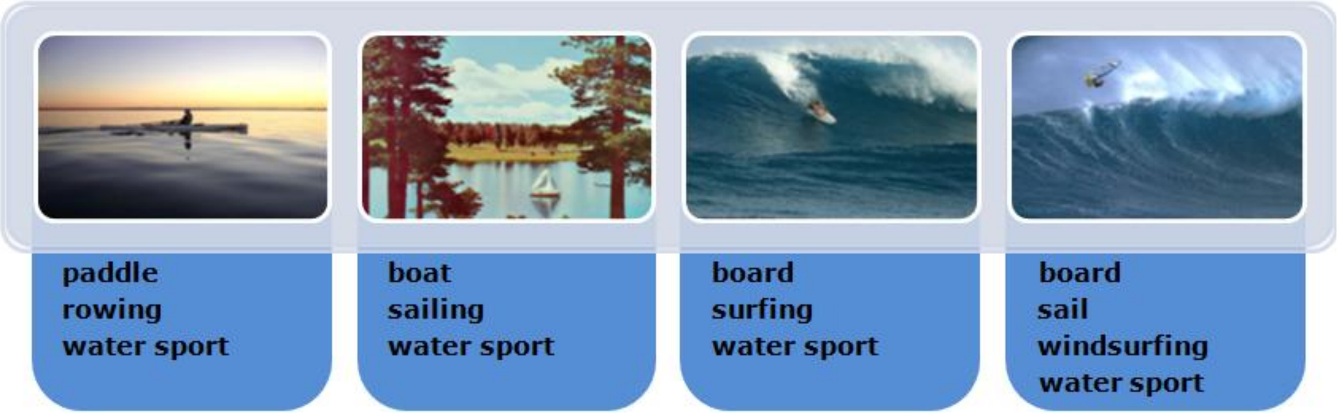


Fig. 7. Examples of semantically annotated images with our automatic system.

Some samples of automatically annotated images with our approach are presented in Fig. 7. More results are discussed in Section III.

III. Results and Evaluation

In order to test our segmentation and labeling approach for the automatic image annotation application, we have built a database called 'water sport image dataset' based on two standards datasets, namely, Berkeley Image dataset and Vitterbi USC-SIPI image database that we have merged and enhanced with Google-retrieved images in order to obtain a broad domain of images suitable for public applications involving image annotation. Berkeley Image dataset contains images in jpeg format with a resolution of 321x481, while Vitterbi USC-SIPI image database is a collection of digitized images in tiff format with an average size of 512x512.

Hence, the 'water sport image dataset' groups together 3148 images of 4 types of common outdoor water sports, namely, 'rowing', 'sailing', 'surfing', and 'windsurfing', with a resolution ranging from 320x433 pixels to 1280x650 pixels and in different image format such as tiff, jpeg, or png. Each category contains from 600 to 800 images. Some samples of our database are displayed in Fig. 4.

This dataset presents challenges of scale, pose and light variations as well as shadow effect and noise due to the water element. In this way, the images of our database have different size and resolution as well as large inter-class similarities, e.g. both windsurfing and surfing sports involve the use of a board, and intra-class variations, e.g. the water color could vary from light blue (Fig. 4 (a)) to dark blue (Figs. 4 (b)-(c)) or even be white (Fig. 4 (d)). Hence, the difficulty of the image segmentation, classification, and annotation in this dataset is very high.

All the experiments have been performed on a commercial computer with a processor Intel(R) Core(TM)2 Duo CPU T9300 2.50 GHz, 2 Gb RAM and using MatLab (Mathworks, Inc.) software.

To assess the accuracy of our AIA system, we adopt the standard criterion as follows:

$$accuracy = \frac{TP+TN}{TP+TN+FP+FN}, \qquad (4)$$

with *TP*, true positive, *TN*, true negative, *FP*, false positive, and *FN*, false negative.

In the first carried-out experiment, we aim to assess the importance of the precise and semantically meaningful segmentation of the image on the resulting semantic annotation of the image. Thus, the images are segmented using different approaches as presented in Fig. 2. We can observe that if the image is segmented using N-cut or edge detector techniques (Figs. 2 (a)-(b)), it results in semantically incoherent foreground objects. These resulting, meaningless visual information prevent the labeling system to process properly. Hence, a bad segmentation of the image leads to the misclassification of this image, and thus to its incorrect annotation. In the opposite case, when applying our multi-layer multi-feature active contour approach (Figs. 2 (c)-(d)), the segmentation is accurate and provides semantically meaningful foregrounds such as illustrated in Figs. 5 (a)-(d). In this case, the image labeling is performed well, leading to the hierarchical categorization of the images, with a computational time in the range of few ms (Figs. 7 (a)-(d)).

In the second experiment, we assess the influence of the number of layers on the classification accuracy. Images in all the dataset are first segmented by first-layer active contours, and in a second batch, by second-layer multi-layer active contours. The resulting confusion matrices are presented in Fig. 6 (a) and Fig. 6 (b), respectively. It results that more layers have the active contours, better is the image classification, thus the precision of the image annotation. Indeed, at each layer, visual and semantic information are gathered in smaller and more meaningful regions. Thus, the extracted features could be more precisely mapped into linguistic notions, based on which decisions are made, leading to more reliable annotations.

The mean average accuracy reached by our AIA system is 95%. Compared to other approaches, little are automatic in the literature. We can note that [15] achieves 73% of accuracy, however it uses very distant categories. On the other hand, the technique presented in [11] is 84% accurate, but it involves constraining assumptions, e.g. only the center of the image is studied, and thus it is not processing data with foregrounds not in the middle of images. Hence, performance of our fully automatic image annotation method are better than those of the state-of-the-art ones, while our dataset is challenging as it contains closely-related classes and foregrounds not in the center of the images (Figs. 5 (b),(d)) and distracted by noise and/or shadows caused by the water element (Figs. 5 (a)-(d)).

IV. Conclusions

In this paper, we propose (a) original multi-layer active contours segmenting the image into semantically meaningful objects and sub-objects and (b) new unsupervised semantic labeling technique based on trained decision trees relying on both numeric and linguistic concepts. Thus, the novel fully automatic image annotation method based on (a) and (b) is performed by using semantic knowledge and visual content analysis together and is efficient in terms of precision, while being compatible with online applications.

References

[1] T. Alqaisi, D. Gledhill, and J. I. Olszewska, “Embedded double matching of local descriptors for a fast automatic recognition of real-world objects”, in *Proceedings of the IEEE International Conference on Image Processing*, October 2012, pp. 2385-2388.

[2] E. Borenstein and S. Ullman, “Combined top-down/bottom-up segmentation”, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 30, no. 12, pp. 2109-2125, December 2008.

[3] Y. Boykov and G. Funka-Lea, “Graph cuts and efficient N-D image segmentation”, *International Journal of Computer Vision*, vol. 70, no. 2, pp. 109-131, November 2006.

[4] J. Cai, Z.-J. Zha, Y. Zhao, and Z. Wang, “Evaluation of histogram based interest point detector in web image classification and search”, in *Proceedings of the IEEE International Conference on Multimedia and Expo*, July 2010, pp. 613-618.

[5] Y. Deng and B. S. Manjunath, “Unsupervised segmentation of colour-texture regions in images and video”, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 23, no. 8, pp. 800-810, August 2001.

[6] I. Endres, A. Farhadi, D. Hoiem, and D. A. Forsyth, “The benefits and challenges of collecting richer object annotations”, in *Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition Workshops*, June 2010, pp. 1-8.

[7] S. Gould, R. Fulton, and D. Koller, “Decomposing a scene into geometric and semantically consistent regions”, in *Proceedings of the IEEE International Conference on Computer Vision*, September 2009, pp. 1-8.

[8] Y.-F. Huang and H.-Y. Lu, “Automatic image annotation using multi-object identification”, in *Proceedings of the IEEE Pacific-Rim Symposium on Image and Video Technology*, November 2010, pp. 386-392.

[9] D. K. Iakovidis and C. V. Smailis, “Efficient semantically-aware annotation of images”, in *Proceedings of the IEEE International Conference on Imaging Systems and Techniques*, May 2011, pp. 146-149.

[10] M. Kass, A. Witkin, and D. Terzopoulos, “Snakes: Active contour models”, *International Journal of Computer Vision*, vol. 1, no. 4, pp. 321-331, January 1988.

[11] S. Kim, S. Park, and M. Kim, “Image classification into object/non-object classes”, in *Proceedings of the International Conference on Image and Video Retrieval*, July 2004, pp. 393-400.

[12] P. Kohli, L. Ladicky, and P. Torr, “Robust higher order potentials for enforcing label consistency”, *International Journal of Computer Vision*, vol. 82, no. 3, pp. 302-324, May 2009.

[13] I. Kokkinos and P. Maragos, “An expectation maximization approach to the synergy between image segmentation and object categorization”, in *Proceedings of the IEEE International Conference on Computer Vision*, October 2005, pp. I.617-I.624.

[14] A. Levin and Y. Weiss, “Learning to combine bottom-up and top-down segmentation”, *International Journal of Computer Vision*, vol. 81, no. 1, pp. 105-118, January 2009.

[15] Y. Liu, D. Zhang, and G. Lu, “Region-based image retrieval with high-level semantics using decision tree learning”, *Pattern Recognition*, vol. 41, no. 8, pp. 2554-2570, August 2008.

[16] A. Mojsilovic, J. Gomes Boykov, and B. Rogowitz, “Semantic-friendly indexing and querying of images based on the extraction of the objective semantic cues”, *International Journal of Computer Vision*, vol. 56, no. 1-2, pp. 79-107, January 2004.

[17] G. S. Muralidhar, A. C. Bovik, J. D. Giese, M. P. Sampat, G. J. Whitman, T. Miner Haygood, T. W. Stephens, and M. K. Markey, “Snakules: A model-based active contour algorithm for the annotation of spicules on mammography”, *IEEE Transactions on Medical Imaging*, vol. 29, no. 10, pp. 1768-1780, October 2010.

[18] S. Nowozin, C. Rother, S. Bagon, T. Sharp, B. Yao, and P. Kohli, “Decision tree fields”, in *Proceedings of the IEEE International Conference on Computer Vision*, November 2011, pp. 1668-1675.

[19] J. I. Olszewska, Unified Framework for Multi-Feature Active Contours, PhD Thesis, UCL, 2009.

[20] J. I. Olszewska, “Spatio-temporal visual ontology”, in *Proceedings of the EPSRC/BMVA Workshop on Vision and Language*, September 2011.

[21] J. I. Olszewska, “A new approach for automatic object labeling”, in *Proceedings of the EPSRC/BMVA Workshop on Vision and Language*, December 2012.

[22] J. I. Olszewska, “Multi-target parametric active contours to support ontological domain representation”, in *Proceedings of the RFIA Conference*, January 2012, pp. 779-784.

[23] J. I. Olszewska and T. L. McCluskey, “Ontology-coupled active contours for dynamic video scene understanding”, in *Proceedings of the IEEE International Conference on Intelligent Engineering Systems*, June 2011, pp. 369-374.

[24] J. I. Olszewska and D. Wilson, “Hausdorff-distance enhanced matching of scale invariant feature transform descriptors in context of image querying”, in *Proceedings of the IEEE International Conference on Intelligent Engineering Systems*, June 2012, pp. 91-96.

[25] S. B. Park, J.W. Lee, and S. K. Kim, “Content-based image classification using a neural network”, *Pattern Recognition Letters*, vol. 25, no. 3, pp. 287-300, February 2004.

[26] J. R. Quinlan, “Induction of decision trees”, *Machine Learning*, vol. 1, no. 1, pp. 81-106, March 1986.

[27] B. C. Russell, A. Torralba, K. P. Murphy, and W. T. Freeman, “LabelMe: A database and web-based tool for image annotation”, *International Journal of Computer Vision*, vol. 77, no. 1-3, pp. 157-173, May 2008.

[28] J. Shao, D. He, and Q. Yang, “Multi-semantic scene classification based on region of interest”, in *Proceedings of the IEEE International Conference on Computational Intelligence for Modelling, Control and Automation*, December 2008, pp. 732-737.

[29] J. Shi and J. Malik, “Normalized cuts and image segmentation”, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 22, no. 8, pp. 888-905, August 2000.

[30] J. Shotton, J. Winn, C. Rother, and A. Criminisi, “TextonBoost for image understanding: Multi-class object recognition and segmentation by jointly modelling texture, layout, and context”, *International Journal of Computer Vision*, vol. 81, no. 1, pp. 2-23, January 2009.

[31] M. Tabb and N. Ahuja, “Multiscale image segmentation by integrated edge and region detection”, *IEEE Transactions on Image Processing*, vol. 6, no. 5, pp. 642-655, May 1997.

[32] D. Zhang, M. M. Islam, and G. Lu, “A review on automatic image annotation techniques”, *Pattern Recognition*, vol. 45, no. 1, pp. 346-362, January 2012.

Representation Modeling Persona by using Ontologies: Vocabulary Persona

GAOU Salma [1]
1.Faculté des sciences . Université
Abdelmalek Essaadi.
Av Jbel Hayane Lot Oulianti N 7
ETG 3, BP 565. Tétouan,MAROC.

EL KADIRI Kamal Eddine[2]
2. Faculté des sciences de Tétouan.
Université Abdelmalek Essaadi.
Mhannech II, B.P : 2121.
Tétouan,MAROC.

CORNELIU BURAGA Sabin [3]
3.Faculty of Computer Science. "A.
I. Cuza" University.
16, G-ral Berthelot Street. Iasi,
700483 ROMANIA.

***Abstract*—Semantic Web is then to add to all these resources semantics that allow computer systems to "understand" the meaning by accessing structured collections of information and inference rules that can be used to drive reasoning automated to better satisfy user requirements. Standard description of Web resources proposed by the W3C, as the name implies, RDF (Resource Description Framework) is a meta-data used to guide the description of resources, to make it more "structured" information necessary for engines research and, more generally, to all necessary computer automated tool for analyzing web pages. The web is a new web sematique or all Web resources are described by metadata, which allows machines better use of these resources. Considering as a foundation specification FOAF (Friend Of A Friend), we use semantic structures (RDFa) to create an ontology and technologies in which it is implemented.Create a conceptual model (eg, an ontology) for personas and their uses in the context of human-computer interaction we will present some screenshots of execution of application.**

Keywords—Semantic Web; FOAF; Persona; Vocabulary; Ontology; Persona;

I. Introduction

Semantic Web technologies and Semantic Web offers us a new approach to managing information and processes, the fundamental principle is the creation and use of semantic metadata. Using the semantics, we can improve the way information is presented. At its simplest, instead of providing a linear search results list, the results can be grouped by meaning. The use of semantic metadata is also crucial for the integration of information from heterogeneous sources, either within an organization or across organizations. [1]

For their implementation, management information, the integration of effective information and intergration of applications require that all information and underlying be semantically described and managed process, is that they are associated with a machine description of their meaning. This, the basic idea behind the Semantic Web has become very important to late 1990s [2] and in a more developed in the 2000s [3] form. The last half decade has seen intense activity in the development of these ideas, especially under the auspices of the World Wide Web Consortium (W3C) [4]. At the heart of all Semantic Web applications is the use of ontologies. A commonly accepted definition of an ontology is: "An ontology is an explicit and formal specification of a conceptualization of a domain of interest" [5] This definition focuses on two points, first, the conceptualization is formal and therefore allows reasoning by. the computer, and the ontology is designed to perform a particular area of interest. Ontology consists of concepts (also known as the class name), relationships (properties) and instances and axioms. Therefore, a brief definition was proposed as <C,R,I,A> 4-tuple, where C is a set of concepts, R a set of relations, I have a set of instances and A a set of axioms [6]. The first work in Europe and the United States on the definition of ontology languages has converged under the W3C to produce an OWL [7] Web Ontology Language. OWL provides mechanisms for the creation of all the elements of ontology concepts, instances, properties (or relations) and axioms [1]. OWL is based on the Resource Description Framework (RDF) [8], which is essentially a language for data modelling, also defined by the W3C. RDF is based on the graphs, but usually serialized as XML. Essentially, it's triplets: subject, predicate, object [1]. The Semantic Web is simply a web of data described and linked in order to establish the context or semantics that adhere to the defined grammar and language constructs [9]. In this article we try to create a conceptual model (eg, an ontology) for personas and their use in the context of human-computer interaction. These models are widely used in the design, development and science. They are powerful tools to represent the structures and relationships in order to better understand the complex, talk, or use. Without models, we have to make sense of raw unstructured data, without the benefit of an organizing principle. Good models emphasize the main characteristics of structures and relationships they represent and de-emphasize less important details [10].

II. Ontologies

Originally, ontology is a philosophical concept, in which philosophers have attempted to account for the existence of formal way. Semantic Web researchers have adapted this term in their own jargon and various definitions of ontology exist in computer literature.

However, the definition most referenced and also the plastic is that of [11] : << ontology is an explicit formal specification of a conceptualization. >> [14] emphasizes safe conceptual nature of an ontology made in order to share knowledge between humans and systems, and between systems. He considers conceptualization, sharing and reuse as the key concepts of an ontology. [15] defines an ontology as an explicit formal description of concepts in a domain of discourse (classes sometimes called concepts), properties of each concept describing the characteristics and attributes of the concept (attributes sometimes called roles or properties) and restrictions on the attributes (sometimes called facets restriction roles). Class concept and define the same term in what follows. A methodology for the development of an ontology is developed in [15].

In summary, computer science, an ontology is understood as a system of basic concepts that are related with each other and represented in a form understandable by a computer each.

The semantic web uses ontologies for various reasons. She uses a simple way to improve search relevance, the request expressed can only refer to a precise concept instead of using keywords ambiguous. More advanced applications using ontologies to combine information from a page of knowledge structures and inference rules.

Many computer languages have emerged to build and manipulate ontologies. In order to develop a standardized language, the W3C has created the web group that has developed OWL. The acronym OWL includes both OIL specifications developed by the European community and Damil language developed by DARPA. OIL is used to define ontologies and DAML adds a few features to RDF Schema to make it easier to define new languages for communication between intelligent agents. OWL therefore defines a syntax for describing RDF vocabularies and build to create ontologies. However, OWL has three sub languages of increasing phrase designed for communities of developers and users that are specific:

- OWL Lite: is the OWL language in the easiest, it is intended to represent hierarchies of simple concepts.
- OWL DL: is more complex than the previous one, it is based on description logic as its name (OWL Description Logics). It is suitable for reasoning, and it guarantees the completeness and decidability of reasoning.
- OWL Full: is the most complex version of OWL, intended for situations when it is important to have a high level of description capability, even if they can not guarantee the completeness and decidability of calculations related to the ontology [12].

OWL ontologies are in the form of text files, documents OWL. The creation of an OWL document is subject to various recommendations [16].

Ontologies are formal and consensual specifications of conceptualizations that provide a shared understanding of a domain, an understanding that can be communicated across people and application systems [13]. This section shows the main components of an OWL ontology

Representation of semantics is necessary for the integration of systems, techniques that knowledge representation is used, it examines how to transform the expression of meaning in a formal representation manipulable by a machine, one way the most used is the ontology.

In this section, and based on [17] and [18] we will explain in detail each one of the above plus two components namely the axioms and instances:

- Concepts: A concept can be a physical object, a concept, an idea [19], the concepts are also called terms or class of ontology, are the objects manipulated by the base ontologies. They correspond to the relevant abstractions of the problem domain, selected according to the objectives that is given and the intended use of the ontology application. They are present in OWL by owl: Class.
- Relationship: translate the interactions between the concepts present in the domain analysis. These relationships are formally defined as any subset of a Cartesian product of n sets, that is to say R: C1 x ... x Cn C2x and include specialization relation (subsumption), the relationship of composition (meronymy), the instantiation relation, etc.. These relationships allow us to capture the structure and the interaction between concepts, which can represent a large part of the semantics of the ontology. They are presented in OWL by owl: ObjectProperty.
- Functions: are special cases of relationships in which the nth element of the relationship is uniquely defined from the n-1 preceding elements. Formally, functions are defined as follows: F: C1 C2 x ... x Cn-1 → Cn
- Axioms: can model always true assertions about abstractions in the field resulted in the ontology. They can combine concepts, relations and functions to define rules of inference which can be made, for example, the deduction, the definition of concepts and relationships, and then to restrict the values of the properties or arguments a relationship.
- Instances: or individuals constitute the extensional definition of ontology. They represent unique elements conveying knowledge about the problem domain. They are defined by OWL owl: Thing.

The engineering knowledge (IC) has long been considered the favorite field of development of expertise in system design based on knowledge [11]. Thus, the operating mechanism by inference, a type representation as declarative ontology, while following the inference rules defined in this ontology, is the source of the Intelligence system. The knowledge engineering has given rise to ontological engineering, where the ontology is the key object which must be addressed. The need for ontology and ontological engineering of knowledge-based systems beginning to be understood and accepted [20].OWL ontologies are used to model domain knowledge [9].

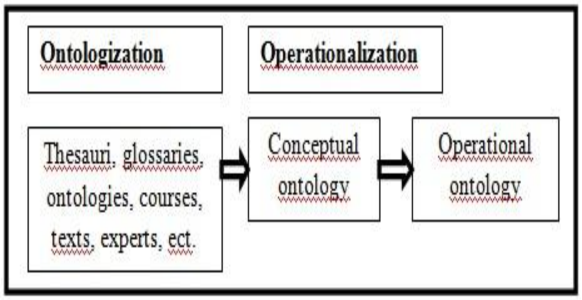


Fig. 1. Etapes de construction d'une ontologie

OWL provides a built-in class whose members correspond to modular parts of a semantic model. It is customary for the URI of an Ontology to correspond to the URL of the file on the Web where the ontology is stored[26]. OWL extends the expressivity of RDFS with additional modeling primitives. For example, OWL defines the primitives owl: equivalentClass and owl :equivalentProperty[27].

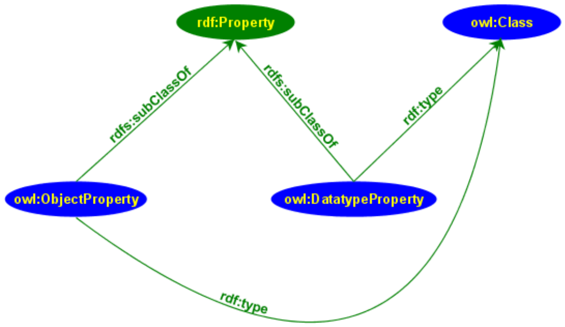


Fig. 2. OWL properties, classes, relations[21]

III. FOAF VOCABULARY AND SYNTAX

FOAF is not so much an application as ontology used by many applications. The Friend of a Friend project (FOAF) was one of the first to recognize the simple power of social networks. The FOAF project provides tools to connect people so a model that contains typical social attributes such as name, email address, interests, etc. [9]. The FOAF project is based on the use of machine readable Web homepages for individuals, groups, businesses and other types of things. For this, we use the "FOAF vocabulary" to provide a set of basic conditions that can be used in these web pages. At the heart of the FOAF project is a set of definitions to use a dictionary of terms that can be used to express claims about the world. The initial objective of FOAF has been on the description of the people, because people are the things that connect most of the other kinds of things that we describe in the Web: they are documents, attend meetings, are shown photos, and so on. FOAF vocabulary definitions presented here are written using a computer language (RDF / OWL) which makes it easy for software to handle some basic facts about the terms of the FOAF vocabulary, and consequently the things described in the FOAF document. FOAF document, unlike a traditional web page, can be combined with other materials to create a FOAF unified database of information [22].

Example[25]

Here is a very basic document describing a person:

```
<foaf:Person>
  <foaf:name>Dan Brickley</foaf:name>

<foaf:mbox_sha1sum>241021fb0e6289f92815fc210f9e9137262c252e</foaf:
mbox_sha1sum>
  <foaf:homepage rdf:resource="http://rdfweb.org/people/danbri/" />
  <foaf:img    rdf:resource="http://rdfweb.org/people/danbri/mugshot/danbri-
small.jpeg" />
</foaf:Person>
```

FOAF is an application of the Resource Description Framework (RDF) because the area we describe - people - has so many competing needs a standalone size could not do them any justice. Using RDF, FOAF wins a powerful extensibility mechanism, allowing FOAF based descriptions can be mixed with claims made in any other RDF vocabulary [23].

FOAF, like the Internet itself, is an information system related. It is built using the Semantic Web technology decentralized, and was designed to allow the integration of data across a variety of applications, Web sites and services, and software systems. To achieve this, FOAF adopted a liberal approach to data exchange. It does not require you to say anything about yourself or others, or puts no limits on the things you say or the variety of the Semantic Web vocabularies that can be used to do this. The current specification provides a "dictionary" basic term about people and the things they do and do [24].

IV. PERSONA

A persona is a typical user (the famous archetype), a fictional representation of target users, which can be used to set priorities and guide our design decisions interface [27].

The method is a technique personas Users centered design, initiated by Alan Cooper in 1999. This method can provide a common and shared vision of the users of a service or product, highlighting their goals, expectations and potential brakes, and offering a more engaging format. In the field of web persona is a fictional character who represents a targeted group. When designing a website, it may be necessary to define multiple personas that will represent each type of potential visitors. A good persona is not to stereotype users but to create users that seems real. That is why we have set goals and personality traits realistic. Based on the objectives of the persona and its specific characteristics (identity, age, familiarity with computers ...) you should check that the user interface to meet the needs of users represented by the personas [27].

Personas give us a precise way of thinking and communicating how users behave, how they think, what they want to accomplish and why. Personas are not real people, but they are based on the behaviors and motivations of real people that we have observed and represent them throughout the design process. They are composite archetypes based on behavioral data collected from many actual users encountered in the ethnographic interviews. Personas are based on patterns of behavior that we see in the research phase, so that we formalize in the modeling phase. Using personas, we can develop an understanding of the goals of our users in specific contexts - an essential tool for the use of user research to inform and justify our designs[28].

Personas are a model used to describe the objectives, skills, abilities, experience and technical context of the users. They are detailed descriptions of archetypal users built on understanding; very specific data models on real people. A character is not based on an individual - he is a construct developed through a detailed process, not the result of a search for the "right" (see the character creation for more details) . They are used by the design team (and largest project team) to describe and keep the foreground user (s) for which the system will be built [27].

Personas, like many powerful tools are a simple concept but must be applied with considerable sophistication. It is not enough to whip up a couple of profiles based on stereotypes and generalizations users; it is not particularly useful to include a photograph of a stock job title and call it a "persona." Personas to be effective tools for design, considerable rigor and finesse should be applied to the process of identification of significant and meaningful in user behavior trends and transform these into archetypes that represent a wide range of users [28].

V. CATEGORIES AND PERSONAS CONSTRUCTION

Persona is a technical approach to ensure the inclusion and optimizing the user experience in the design of an interactive medium. To be useful, the persona must come from real information about users without their creation which may be based on stereotypes. This approach allows one hand to filter and synthesize data users and secondly to unite all stakeholders around key profiles: the main tasks that should be an answer, user's needs and priorities appear more easily identifiable.

Categories of persona:

- Primary Persona

This persona is usually designated the primary persona. Indeed, each primary persona requires the presence of its own user interface in a particular application. Knowing that there will be more of a primary persona when their needs cannot be met by the same interface. The fewer the number of primary personas the better.

- Secondary Persona.

By focusing on the primary persona, the secondary persona's goals and needs can mostly be met. Nevertheless, there are a few needs specific to them that are not a priority for the primary persona. To meet the needs of a secondary persona, there may be small additions to the interface necessary. However, these additions should not negatively affect the experience of the primary persona.

- Supplemental persona

User persona which are neither primary nor secondary are called supplemental persona. The combination of primary and secondary personas represents completely supplemental persona's needs that are completely satisfied by the solution devised for one of our primaries.

- Customer persona

Customer personas match customer needs, as discussed by ((auteur) et al., (année)) and their treatment is similar to that of secondary personas.

- Served persona

Served persona some what differ from persona types discussed previously. Although they are directly affected by the use of the product, they are not users of the product at all.

- Negative persona

Negative persona they aren't users of the product, like served personas. They mediate between stakeholders and product team members by informing them that there are specific types of users that the product is not being built to serve [28].

The majority of studies that have been done on personas seem to focus on targets in the context of what distinguishes persona to another [29]. But with narrative perspective objectives are parts of what makes the persona act in a given situation. What difference personas are like in real life, personal characteristics possess persona (age, history, psyche, etc.).

Personas Construction

As a rounded character [30], the persona can be characterized by the following elements, namely:

- Body: body constitutes a human being. Sex, age, look helps the designer emphasise the Persona
- Psyche: to understand motivations for actions we need to understand what lies behind the motivation, the personality.
- Background : job position, family, education, social- and cultural positions explain motivations for actions.
- Emotional state : to know the emotional state furthers engagement in the Persona [31]. Inner needs and goals, ambitions and wishes create a foundation for the emotional state.
- Cacophony : two oppositional character traits [32]. The oppositional traits are what constitute the difference between a stereotype and a rounded character.

Persona is static but is dynamic when inserted into the actions of the scenario. In the scenario, the persona is in a context, in a specific situation with a specific purpose [33].

Persona Elements:

- Goals ,
- Attitudes (related to your context) ,
- Behaviors & Tasks (in your context) ,
- Name ,
- Photo ,
- Tagline ,
- Demographic Info (brief just to help "humanize" them) ,

- Skill level ,
- Environment ,
- Scenarios (not all but perhaps the highest priority, most common or most telling about their needs) [34].

VI. The Persona Data Model And The Diagram

As you can see, the persona model focuses on six categories: context, person, goals, image, organization and event.

The context class is extended by four sub-classes of society, the application of the contact points, which is linked to the address and location (country). Persona of each person must be identified by name, age and other demographic information. He has goals that are related to the application, and includes individuals, businesses and concrete targets. Each persona must have an image that is particularly a photograph that corresponds with the name. It should have affilation organization that provides knowledge and experience, are included education, training and expertise. This should not be limited only to the application. Persona must attend the event.

A. *Use case diagram*

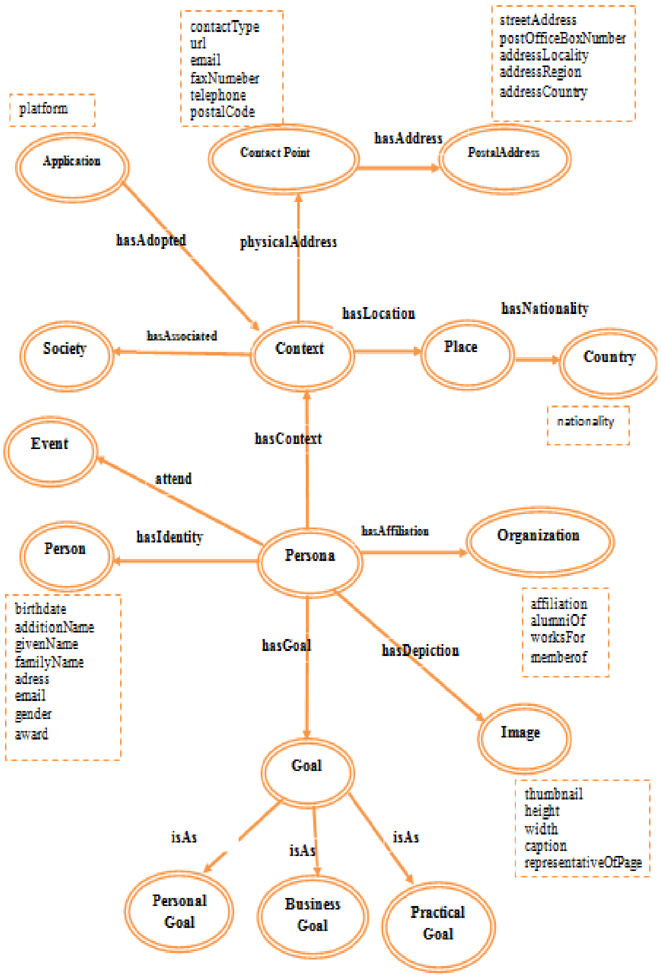


Fig. 3. Use case diagram of the Persona data model[35].

B. *Class diagram*

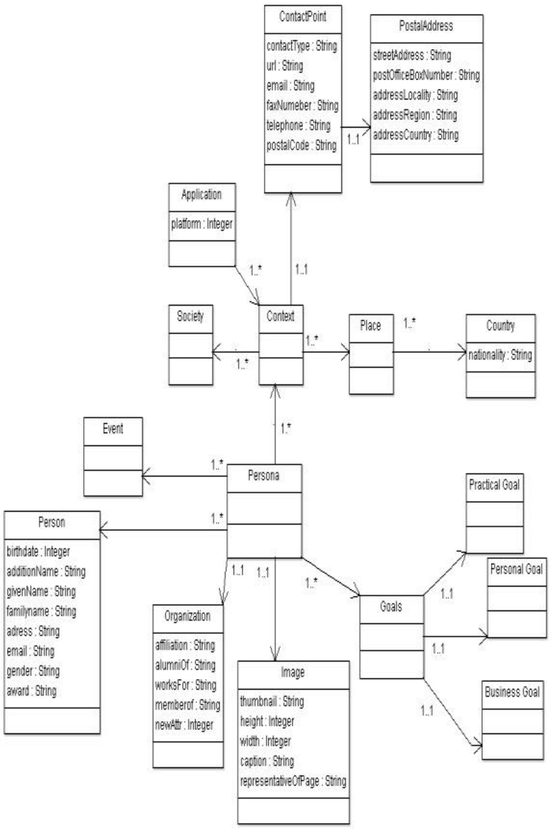


Fig. 4. Class diagram of the Persona data model.

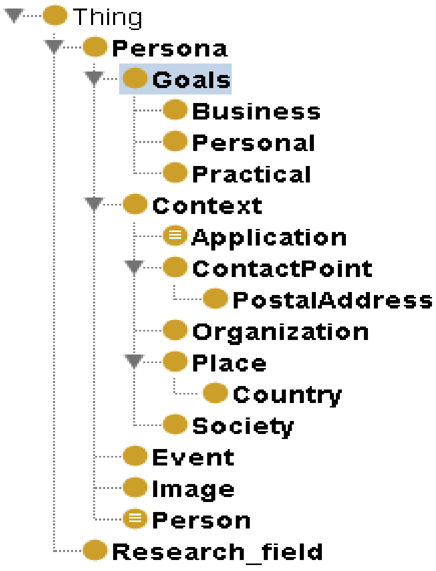
TABLE I. THE TABLE BELOW SUMMARIZES ALL CLASS RELATIONS IN PERSONA.

Subject Class	Object Class	Definition
Persona	**Person**	indicates that persona is identified by demographic information for person
Persona	**Context**	Indicates that **Persona** has Context a **context**
Persona	**Goals**	Every **Persona** should have **Goals**
Persona	**Image**	Every **Persona** should have **Image**
Persona	**Organization**	**Persona** has affiliated to **Organizations**
Persona	**Event**	**Persona** should attend the **Event**
Context	**Place**	**Place** of the location.
Context	**ContactPoint**	Indicates that **Context** has a **Contact point**
Application	**Context**	**Application** has adopted **Context**
Context	**Society**	**Context** has associated to **Society**
Contact Point	**PostalAddress**	**ContactPoint** has address in

		PostalAddress
Place	**Country**	**Place** has associated to **Country** by nationality
Goal	**Practical Goal**	Indicates that **Goal** are **practical Goal** efficiently
Goal	**Personal Goal**	**Personal Goal** should get an adequate amount of work done, be comfortable and fun
Goal	**Business Goal**	quality education is required for **Business Gola**, for exemple

VII. Protege

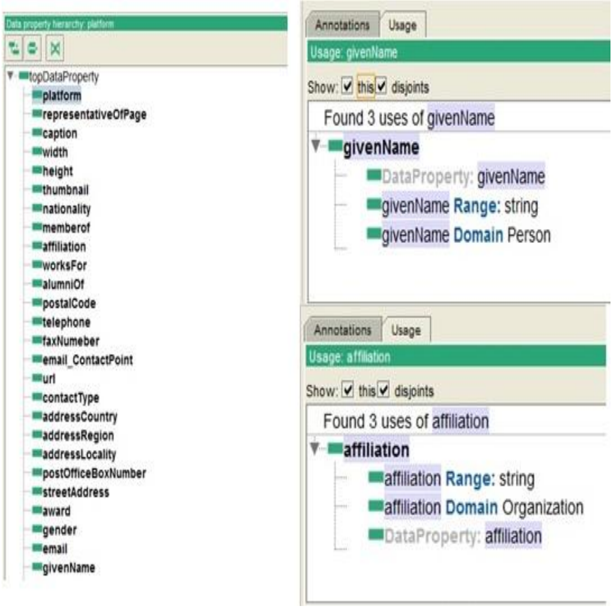
As part of our modeling tool we have chosen for the creation of ontology is none other than Protege2000 (Protégé200, 2008) because it allows viewing and comfortable and intuitive manipulation of concepts and relations up the ontology. Open source ontology editor, developed at the Department of Medical Informatics at Stanford University. The ontology editor "Protege" was used to edit the ontology of Persona with the aim to automatically generate the code for OWL, as well as for generating HTML documentation for our ontology. It should be noted that "Protege" offers very sure a lot of features, and we certainly did not all used. Protege also program or import a large number of "plugins" that can be downloaded directly from its official website. These plugins provide new features such as the ability to edit the ontology with different formats for describing ontology (RDF (S), OIL, DAML + OIL, OWL).



```
Thing
  Persona
    Goals
      Business
      Personal
      Practical
    Context
      Application
      ContactPoint
        PostalAddress
      Organization
      Place
        Country
      Society
    Event
    Image
    Person
  Research_field
```

Fig. 5. Gives the list of classes in a hierarchical view.

A. *This is the list of Data properties:*



```
topDataProperty
  platform
  representativeOfPage
  caption
  width
  height
  thumbnail
  nationality
  memberof
  affiliation
  worksFor
  alumniOf
  postalCode
  telephone
  faxNumeber
  email_ContactPoint
  url
  contactType
  addressCountry
  addressRegion
  addressLocality
  postOfficeBoxNumber
  streetAddress
  award
  gender
  email
  givenName
```

Usage: givenName

Show: ☑ this ☑ disjoints

Found 3 uses of givenName

```
givenName
  DataProperty: givenName
  givenName Range: string
  givenName Domain Person
```

Usage: affiliation

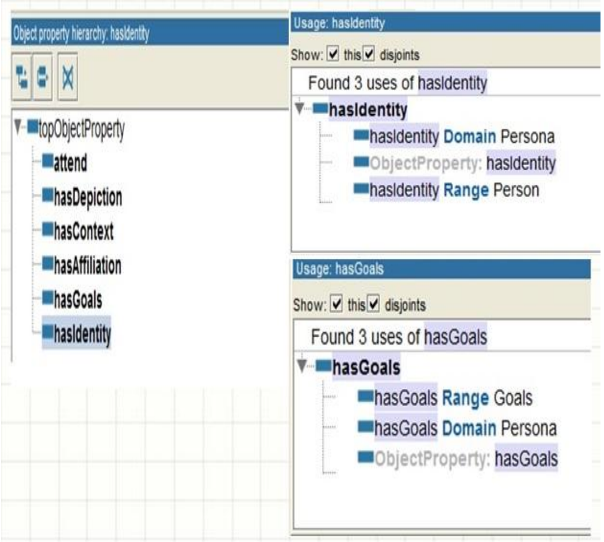
Show: ☑ this ☑ disjoints

Found 3 uses of affiliation

```
affiliation
  affiliation Range: string
  affiliation Domain Organization
  DataProperty: affiliation
```

Fig. 6. Extract the Data property defined in the ontology Persona

B. *This is the list of object properties:*



```
topObjectProperty
  attend
  hasDepiction
  hasContext
  hasAffiliation
  hasGoals
  hasIdentity
```

Usage: hasIdentity

Show: ☑ this ☑ disjoints

Found 3 uses of hasIdentity

```
hasIdentity
  hasIdentity Domain Persona
  ObjectProperty: hasIdentity
  hasIdentity Range Person
```

Usage: hasGoals

Show: ☑ this ☑ disjoints

Found 3 uses of hasGoals

```
hasGoals
  hasGoals Range Goals
  hasGoals Domain Persona
  ObjectProperty: hasGoals
```

Fig. 7. Extract the Object property defined in the ontology Persona

VIII. Strengths Of Personas As A Design Tool

- Persona is a powerful design tool, multi-purpose, which helps to overcome several problems that currently plague the development of digital products. Personas help designers:

- Determine what a product should do and how it should behave. Persona goals and tasks form the basis of the design effort.

- Communicate with stakeholders, developers and other designers. Personas provide a common language for discussing design decisions and also help to keep the design centered on users at every stage of the process.

- Build consensus and commitment to the design. With a common language for a common understanding. Personas reduce the need for complex schematic models, it is easier to understand the many nuances of user behavior through the narrative structures that personas employ.

- Measure the effectiveness of the design. Design choices can be tested on personas in the same way that they can be presented to a real user during the training process. Although it does not replace the need to test with real users, it provides a powerful reality-check for designers trying to solve design problems. This allows design iteration to produce quickly and cheaply to the table, and the result is a much stronger base of design when it comes time to test with real people.

- Contribute to other product-related efforts such as marketing and sales plans. The authors have seen their customers repurpose personas within their organization, informing marketing campaigns, organizational structure, and other strategic planning activities. Business units outside of product development desire thorough knowledge of users of a product and see generally personas with great interest[33].

- You can use personas when you want:

- Making sense of research findings:To do this, you need to analyze the results and identify trends in research and capture the most important information about who they are, what they need to accomplish their skills, abilities and pain points, etc.

- Plan your product: To do this, analyze the competition through the eyes of your personas, features brainstorming possible using your personas and persona characteristics using a weighting matrix based on priorities.

- Explore design solutions: You need to prepare scenarios, mapping of the design, mood boards and explorations of visual design.

- Assess your solutions: You must make cognitive walkthroughs, design reviews with personas, user testing, user research underway with personas, Quality Assurance (QA) testing and bug bashes (focus quality assurance tests and create persona based on test cases, bug persona labelling.

- Supporting the release of your product: You will need to do the documentation, training, support materials (personas can help focus the teaching materials, guides, and editorial content), marketing and sales (tailor efforts based on personas, the distinction between users and customers (students compared, for example))

- Communicate with the project team and beyond: You have to share your learning with the rest of the team, gain consensus for designing early on ... before conception occurs, hang your personas around the room to keep the project development and create a common language and a shared vision [36].

IX. Deploying The Application Persona

Persona web application developed using JSP and Servlet technologies. To be functional, it must be deployed in an HTTP application server with JSP / Servlet container. Apache Tomcat is both HTTP server (Apache) and Servlet / JSP container, which makes it an ideal candidate for the deployment of our application. The figure below shows the deployment scheme of Persona application.

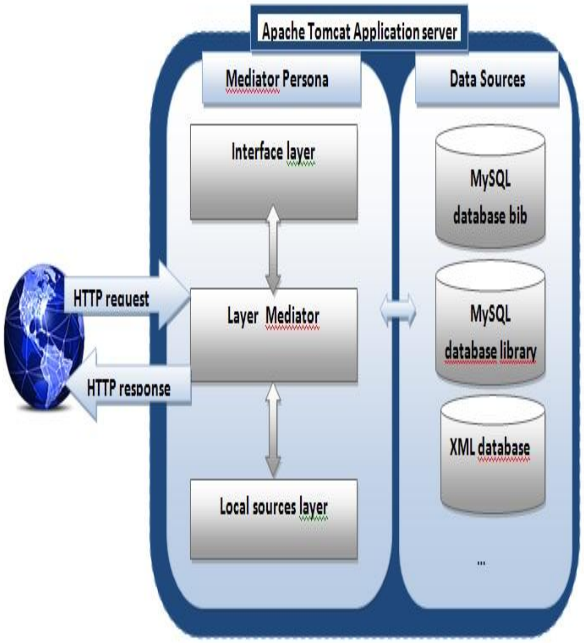


Fig. 8. Diagram of deployment Persona Application

The deployment of the application on Apache Tomcat persona is done by placing the directory containing the application files in the webapps directory of Tomcat.

X. Images Of Application Execution

In what follows we will present some screenshots of execution of our application, the homepage is the first window website, where you can access the different menus of the site persona. It contains five main menus (Identity, Organization, Context, Event, and Goals).

- Context menu contains three submenus (Place, Society and Contact point).

- Goals menu contains three submenus (Personal, Business and Practical).

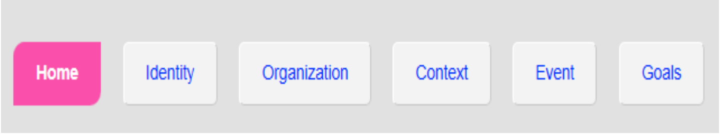


Fig. 9. Menus of the site persona.

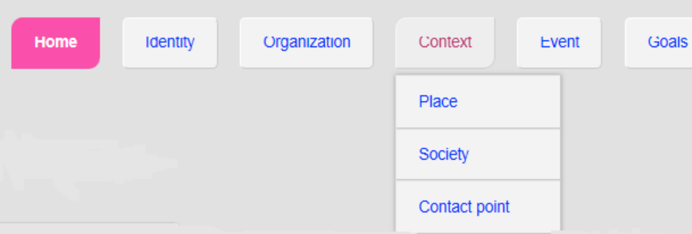
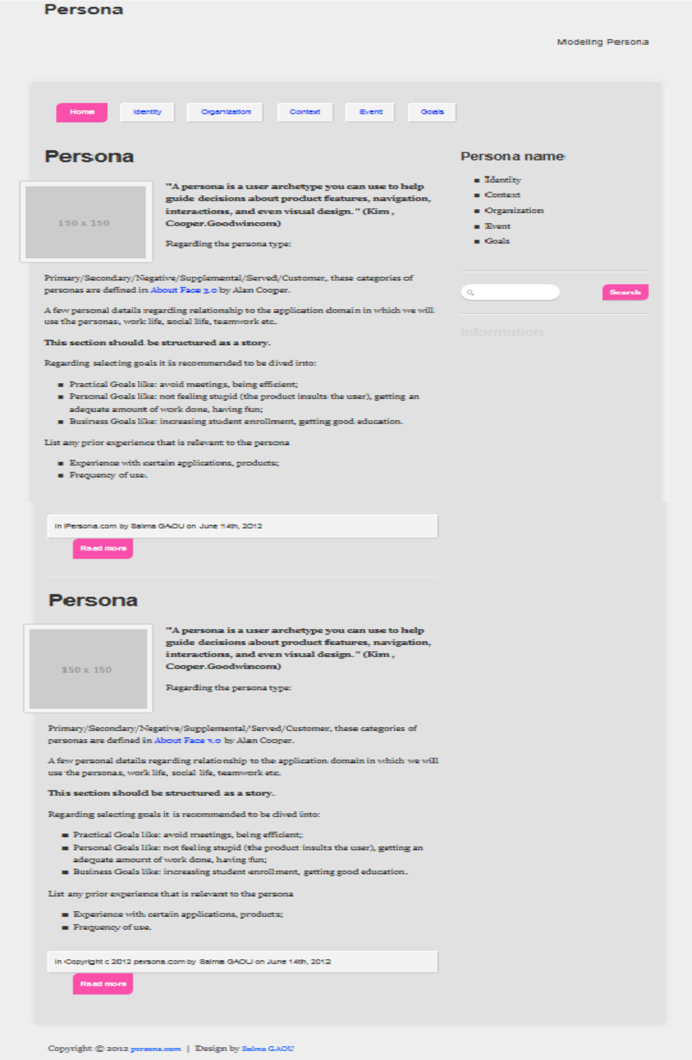


Fig. 10. Context menus active with submenus.



Persona

"A persona is a user archetype you can use to help guide decisions about product features, navigation, interactions, and even visual design." (Kim , Cooper.Goodwincom)

Regarding the persona type:

Primary/Secondary/Negative/Supplemental/Served/Customer, these categories of personas are defined in About Face 3.0 by Alan Cooper.

A few personal details regarding relationship to the application domain in which we will use the personas, work life, social life, teamwork etc.

This section should be structured as a story.

Regarding selecting goals it is recommended to be dived into:

- Practical Goals like: avoid meetings, being efficient;
- Personal Goals like: not feeling stupid (the product insults the user), getting an adequate amount of work done, having fun;
- Business Goals like: increasing student enrollment, getting good education.

List any prior experience that is relevant to the persona

- Experience with certain applications, products;
- Frequency of use.

In iPersona.com by Salma GAOU on June 14th, 2012

Persona name

- Identity
- Context
- Organization
- Event
- Goals

Persona

"A persona is a user archetype you can use to help guide decisions about product features, navigation, interactions, and even visual design." (Kim , Cooper.Goodwincom)

Regarding the persona type:

Primary/Secondary/Negative/Supplemental/'Served/Customer, these categories of personas are defined in About Face 3.0 by Alan Cooper.

A few personal details regarding relationship to the application domain in which we will use the personas, work life, social life, teamwork etc.

This section should be structured as a story.

Regarding selecting goals it is recommended to be dived into:

- Practical Goals like: avoid meetings, being efficient;
- Personal Goals like: not feeling stupid (the product insults the user), getting an adequate amount of work done, having fun;
- Business Goals like: increasing student enrollment, getting good education.

List any prior experience that is relevant to the persona

- Experience with certain applications, products;
- Frequency of use.

In Copyright c 2012 persona.com by Salma GAOU on June 14th, 2012

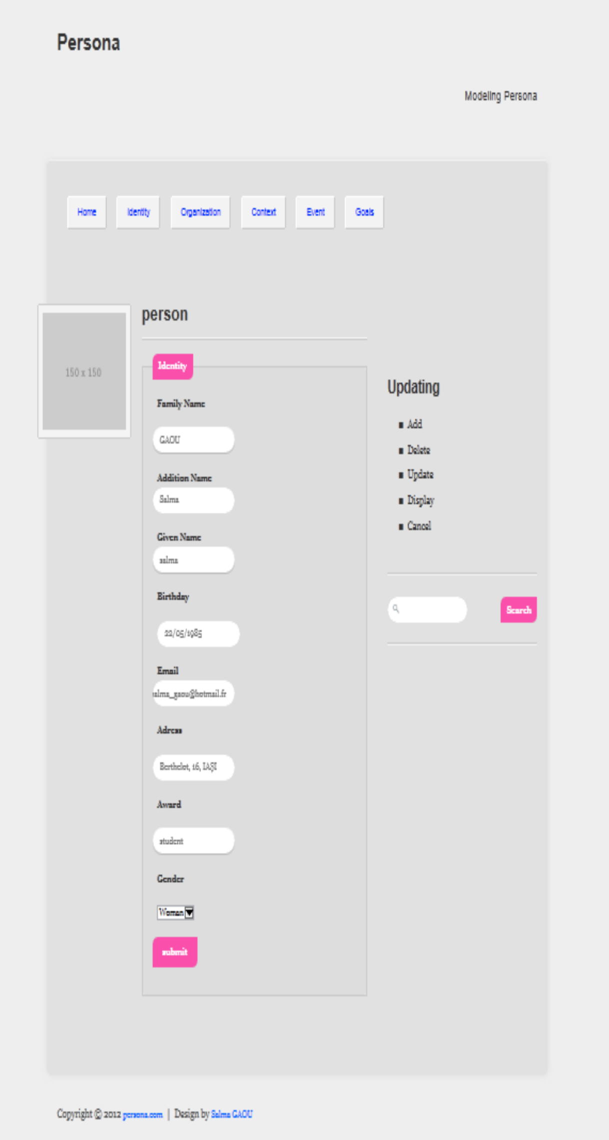
Copyright © 2012 persona.com | Design by Salma GAOU

Fig. 11. the first page of website persona

To access this page below the identity, just click on the menu Identity This file contains all information (family name...) for each persona, also includes many features including:

The addition, modification and removal identity.

The same think for context (place, Society and Contact point), event, goals (Personal, Business and Practical), organization.



person

Family Name: GAOU

Addition Name: Salma

Given Name: salma

Birthday: 22/05/1985

Email: salma_gaou@hotmail.fr

Adress: Berthelot, 16, IASI

Award: student

Gender: Woman

Updating

- Add
- Delete
- Update
- Display
- Cancel

Copyright © 2012 persona.com | Design by Salma GAOU

Fig. 12. page Identity of website persona

XI. Conclusions

In this article, we introduced the notion of ontology and several methods and tools for ontology engineering. In our study, taking into consideration the characteristics and benefits of ontologies, we focused on the ontology for building a persona vocabulary.

Like many powerful tools, personas are a simple concept but must be applied with considerable sophistication. It is not enough for some profiles based on stereotypes and generalizations users, it is not particularly useful to attach photos to a job title and call it a "persona." Personas to be effective tools for design, rigor and considerable finesse should be applied to the process of identification of significant and meaningful in user behavior trends and transform these into archetypes that represent a wide range of users.

In addition, ontologies have the potential to allow a true knowledge sharing and reuse of heterogeneous agents, both human and computer. A major challenge is still open alignment of different ontologies to provide interoperability between heterogeneous agents. Considering that we were able to achieve the objectives of the party in this document, and we made the right choices about the tools for implementation, so that our work will be a great way for other future projects.

Looking ahead, we plan to expand our study using another technique, namely: first, the development of other ontologies and combine with ours to enrich the vocabulary used for annotating and research. Then test the possibility of reasoning provided by OWL. Finally, our next task will be presented the implementation details of the application, and is going to use for the realization of our application: Development Tool: JBuilder7, Virtual Machine, Java: J2SDK 1.4.2, Web Server JSP Tomcat 4.x, Libraries: Jena version 2.3 and Servlet / JSP Version 2.3/1.2 (included in Tomcat 4.x).

ACKNOWLEDGMENTS

Research reported in this paper has been partially financed by doctoral research within the Faculty of Computer Science University, Alexandru Ioan Cuza of Iasi in the program <<Eugen Ionescu >> 2011-2012, and thank you for Dr. Sabin-Corneliu Buraga Associate Professor Faculty of Computer Science in Iasi for helping.

References

[1] Semantic Web Technologies ,trends and research in ontology- based systems,John Davies, BT, UK,Rudi Studer,University of Karlsruhe,Germany,Paul Warren,BT, UK

[2] Berners-Lee T.1999. Weaving the web. Orion Busines Books.

[3] Berners-LeeT , Hendler J, Lassila O.2001.The semantic web. In Scientific American, May 2001.

[4] W3C staff assigned to the W3C Semantic Web Activityare Sandro Hawke (see also private blog), Ivan Herman (see also private blog), and Eric Prud'hommeaux , Ivan Herman, Semantic Web Activity Lead, 2011-11-07. en ligne : http://www.w3.org/2001/sw

[5] Gruber T.1993.A translation approach to portable ontologie. Knowledge Acquisition 5(2) : 199-220, http://Ksl-web.stanford.edu/KSL_Abstracts/KSL-92-71.html

[6] Staab S, Stuber R (Eds). 2004. Handbook on ontologies. International Handbooks on Information Systems. Springer : ISBN 3-540-40834-7.

[7] Ivan Herman 6 September 2007: W3C Launches the OWL Working Group, en ligne : http://www.w3.org/2004/OWL/

[8] 10 February 2004, RDF is a standard model for data interchange on the Web. In ligne : http://www.w3.org/RDF/

[9] Semantic Web Programming,John Hebeler,Matthew Fisher,Ryan Blace,Andrew Perez-Lopez,septembrie 2009

[10] About Face 3 The Essentials of Interaction Design Alan Cooper, Robert Reimann, and Dave Cronin, 2007.

[11] Tim Berners-Lee, James Hendler, Ora Lassila The Semantic Web, Scientific American, May 2001

[12] Thomas R.Gruber. Ontolingua: A mechanism to mobile ontologies. Knowledge Systems Laboratory Technical Report KSL-91-66, Stanford University, version 3.0, CA, 1992.

[13] Xavier Lacot (2005) Introduction à OWL, un langage XML d'ontologies Web.

[14] Personas Added by Gary Thompson, last edited by Daphne Ogle on May 25, 2010.

[15] Riichiro Mizoguchi. The role of ontological engineering in the field of ILE

[16] Natalya F.Noy and Deborah L.McGuinness. Ontology development 101: A guide to creating your first ontology.

[17] Michael k.Smith, Chris Welty, Deborah L.McGuinness. OWL Web Ontology Language-Reference. http://www.w3.org/TR/2004/REC-owl-ref-20040210/ (online June 16, 2005).

[18] Tr.Gruber:. A translation approach to mobile ontology specification. Knowledge Acquisition 5 (2): pp. 199-220, 1993.

[19] Gomez-Perez. Recent developments in matters of design, maintenance and use of ontologies. 3emes meetings TIA Terminology and Artificial Intelligence, Vol 19, pp. 9-20, 2000.

[20] Sticef.org (2003) Apport de l'ingénierie ontologique aux environnements de formation à distance

[21] ATHENA WP4 SKOS Workshop Rome, ICCU, 16-17 July 2009 in ligne : http://www.w3c.it/talks/2009/athena/slides.html#(52)

[22] D-FOAF - SECURITY ASPECTS IN DISTRIBUTED USER MANAGEMENT SYSTEM, Slawomir Grzonkowski, Adam Gzella, Henryk Krawczyk, Sebastian Ryszard Kruk, Francisco J. Martin-Recuerda Moyano, Tomasz Woroniecki , Gdansk University of Technology, ul. Narutowicza 11/12, 80-952 Gdansk, Poland,

[23] Dan Brickley, Libby Miller.The Friend of a Friend (FOAF) project, Edd Dumbill's writings.2000, http://www.foaf-project.org/about.

[24] Dan Brickley, Libby Miller. FOAF Vocabulary Specification 0.98. Namespace Document 9 August 2010 - Marco Polo Edition. Creative Commons Attribution License. http://xmlns.com/foaf/spec/.

[25] Semantic Web for the Working Ontologist Effective Modeling in RDFS and OWL, Second Edition, Dean Allemang,Jim Hendler,2006.

[26] Linked Data Evolving the Web into a Global Data Space,Tom Heath,Christian Bizer, 2011.

[27] PERSONAS: définition et démarche. J.C. GROSJEAN. Quality Street, Coaching Agile, Management Agile & Lean, Expérience Utilisateur, Tests. 29 décembre 2008

[28] About Face 3 The Essentials of Interaction Design Alan Cooper, Robert Reimann, and Dave Cronin, 2007.

[29] Pruitt, J. and J. Grudin (2003). Personas: Theory and Practice. 2003.

[30] Nielsen, L. (2002). From User to Character. DIS2002, London.

[31] Smith, M. (1995). Engaging Characters: fiction, emotion, and the cinema, Clarendon Press.

[32] Horton, A. (1999). Writing the Character-Centered Screenplay. L.A., University of California Press.

[33] Proceedings of the Third Danish Human-Computer Interaction Research Symposium,Morten Hertzum, Computer Science, Roskilde University ,Simon Heilesen, Communication Studies, Roskilde University ,Writings in Computer Science, No. 98 ,Roskilde University, Roskilde, Denmark, 2003.

[34] Persona Creation Added by Daphne Ogle, last edited by Allison Bloodworth on May 26, 2009.

[35] S. Gaou, K. E. El Kadiri, and *S.CORNELIU BURAGA*. Journal International Journal of Computer Science Information and Engineering Technologies l'article << USE OF ONTOLOGIES IN MODELING PERSONA>>.

[36] Personas Added by Gary Thompson, last edited by Daphne Ogle on May 25, 2010.

A Study on the Conception of Generic Fuzzy Expert System for Surveillance

Najar Yousra
Dep. of Informatics,
Higher Institute of Informatics
Tunis, Tunisia

Ketata Raouf
National Institute of Applied
Sciences and Technologies.
Research unit on Intelligent Control
Design and Optimisation of
Complex System (ICOS) Tunisia

Ksouri Mekki
National School of engineering
in Tunis, Tunisia Research Unit
LASC, ENIT

Abstract—**This paper deals with using fuzzy logic to minimize uncertainty effects in surveillance. It studies the conception of an efficient fuzzy expert system that had two characteristics: generic and robust to uncertainties. Analyzing distance between variables optimal and real values is the main idea of the research. Fuzzy inference system decides, then, about significant variables state: normal or abnormal. A comparison between three proposed fuzzy expert systems is presented to highlight the effect of membership number and type. Beside, being generic this system could also be applied in three fields: industrial surveillance, camera surveillance and medical surveillance. To expose results in these fields, matlab is used to realize this approach and to simulate systems responses which revealed interested conclusions.**

Keywords—Generic Fuzzy expert system; surveillance; uncertainty'error analysis ;three tanks; ECG

I. INTRODUCTION

Ambiguous environments constitute an enormous problem for decision makers. As a matter of fact, uncertainties affect decision making especially for surveillance in many fields. These uncertainties are the result of many sources such as: nonlinearities, non exhaustive mathematical models, non effectiveness of sensors/detection equipments and qualitative knowledge representation. The most common methodologies that had dealt with this issue are traditional tools as probability theory, error interval analysis and especially fuzzy theory [1]. The first link between fuzzy theory and decision making was introduced in [2]. It was based on the fact that according to a criterion good solutions are fuzzy sets. Besides, the best solution set is obtained from their intersections [3]. The most popular fuzzy sets approach, in decision-making, is the maximum ranking solutions. This method is natural when interpreting the fuzzy sets as flexible constraints. While uncertainty affects several domains and has many facets (randomness, fuzziness etc.), fields and applications concerned with this issue are, especially in the last decade, growing proving the efficiency of fuzzy logic use.

Fuzzy Expert Systems (FES) are expert system that uses fuzzy sets to reason [5]. In another words, FES are intelligent tools capable of making decisions dealing with ambiguous data. A recent research [4] had proved that, in 2010, that the number of published papers adopting fuzzy systems approaches is the most important. Besides, the same article confirms that industrial and medical applications and especially diagnosis is the most growing application field of these techniques.

II. FUZZY EXPERT SYSTEM OVERVIEW

Fuzzy expert system or fuzzy inference system is composed of three units: fuzzification, inference engine and defuzzification. It treats qualitative data with vague and fuzzy descriptions. The application of fuzzy expert system touched many fields especially industrial and medical surveillance.

Recent works used FES in *fault diagnosis applications*. [11] had realized a diagnosis application which based on FES to identify failures in power system by analyzing amplitude and signal orientation then by classifying abnormalities. In the same spirit, many applications in power system had been developed with FES such as [29][14][8]. Detecting episodes of poor water quality is realized by fuzzy inference system in [15]. In the same domain these works using FES treated aluminum electrolysis [16] and detecting failures in computer [17]. Research results also in developing decision making applications using fuzzy expert systems in *medical diagnosis*. [18] presents a fuzzy application to analyze diabetic state. In another hand, using fuzzy logic many systems take decision: [19] about hypertension state, [20] about lever state, [21] about state of prostate cancer, [22] about heart state, [23] about breast cancer.

TABLE I. FES APPLICATION

FES Application Fields	***Publications about SEF***
Industrial diagnosis	*[45], [30], [31], [32], [33], [28], [34], [35], [36], [37], [38], [39], [40], [41], [42], [6], [43] , [44], [7], [11], [14], [8], [15], [10], [50],[58]*
Medical diagnosis	*[47], [52], [46] ,[48], [49] ,[18], [19], [20],[21],[53],[23], [54], [19], [45]*
Video surveillance	*[54],[60]*
Economic domain	*[55], [56], [51]*
Civil domain	*[34], [10], [27]*
Software domain	*[57]*

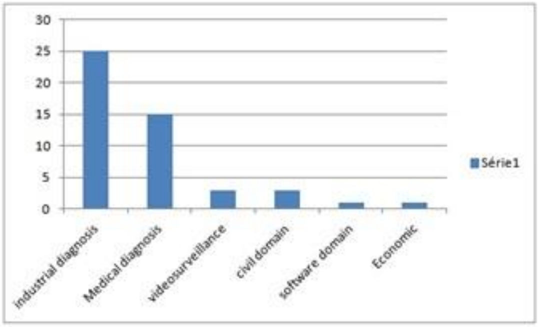


Fig. 1. Fuzzy Expert System publications per domain (50 publications)

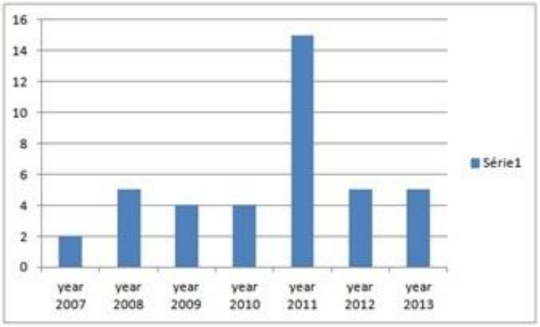


Fig. 2. Fuzzy Expert System publications per year (2007-2013)

This state of art on FES between 2007 and 2013 took into account 50 new publications. We conclude on the importance of fuzzy inference systems in decision making issue. In fact, uncertainty is a matter that affects all kind of field which explains the applications diversification. This research confirms the conclusions made in [4] about the most significant application field which is diagnosis (industrial and medical). We explain these facts by the need of decision aid systems in diagnosis and by the abundance of fuzzy data in these environments. Results generated by different expert systems are robust against vagueness and uncertainty and a certitude coefficient is usually calculated to enhance the effectiveness and the interpretation of outputs.

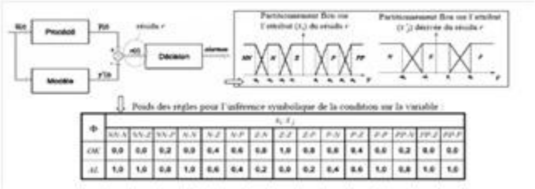
We remark also that each application had its particular inputs and outputs. Thus, the developed fuzzy expert systems are specific for each treated problem. This point had been the key of our research issue.

The abundance of articles in this issue indicated efficient results in industrial and medical diagnosis and surveillance. However, other fields are concerned with fuzzy systems. [52] used genetic algorithm to build rule base, [43] evaluated the performance of software based on certain characteristics, [24] developed FES to evaluate the state of public discharge land, [25] realized a multi agent system and FES decided about the role of each agent, [26] used FES in supply chain localization, [27] used FES in travelling domain, [28] used FES in renewable energy.

III. RESEARCH ISSUE

Our aim is to propose a generic fuzzy expert system that could be applied in several domains to monitor the state of significant variables characterizing the studied situation. In this optic, we should first determinate these variables and fixe their optimal and desired values. Then, the proposed FES is responsible of deciding whether the variable is behaving in optimal trajectory.

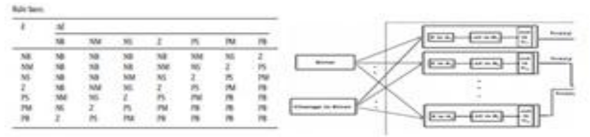
We gave a special attention to research of Evsukoff [7] that presents a FES based on the analysis of significant variables residual and their variation. It was applied in industrial fault detection where partial decisions are made about variables (normal-*OK* or alarm-*AL)*.


Procédé
Modèle
Décision
alarmes
résidu r
Partitionnement flou sur l'attribut (x_i) du résidu r
Partitionnement flou sur l'attribut (x'_i) dérivée du résidu r
NN
N
Z
P
PP
Poids des règles pour l'inférence symbolique de la condition sur la variable :

Φ	NN.N	NN.Z	NN.P	N.N	N.Z	N.P	Z.N	Z.Z	Z.P	P.N	P.Z	P.P	PP.N	PP.Z	PP.P
OK	0,0	0,0	0,2	0,0	0,4	0,6	0,8	1,0	0,8	0,6	0,4	0,0	0,2	0,0	0,0
AL	1,0	1,0	0,8	1,0	0,6	0,4	0,2	0,0	0,2	0,4	0,6	1,0	0,8	1,0	1,0

Fig. 3. Evsukoff fuzzy inference system

In an earlier work [58], we've proposed a modified version of Evsukoff FES which minimized rules number and raised robustness against uncertainties by weighting rules with triangular functions instead of fixed values. In the same way and applied in control, [59] analyzed the error end its variation using fuzzy expert system with seven membership functions for each input. Figure 4 illustrates its inference system.



e	Δe NB	NM	NS	Z	PS	PM	PB
NB	NB	NB	NB	NB	NM	NS	Z
NM	NB	NB	NB	NM	NS	Z	PS
NS	NB	NB	NM	NS	Z	PS	PM
Z	NB	NM	NS	Z	PS	PM	PB
PS	NM	NS	Z	PS	PM	PB	PB
PM	NS	Z	PS	PM	PB	PB	PB
PB	Z	PS	PM	PB	PB	PB	PB

Fig. 4. Panda inference system for error analysis

After studying different points of view, we are trying to determinate the most appropriate fuzzy expert system to be adopted and to be a generic tool for monitor a situation and for helping in decision making about its state.

We are proposing a system which gives partial conclusion about each variable by analyzing its residual. Also, we are studying in this work the effects of raising the number of membership function in FES. Finally, we should prove that the approach is generic by applying same FES in different fields.

IV. CONCEPTION OF FES FOR SURVEILLANCE

Surveillance is, generally, assured through two steps: detection of anomalies and their diagnosis. The FES we are proposing is responsible of detecting abnormal situation. Figure 5 is schema block that defines the inputs and the outputs of the system.

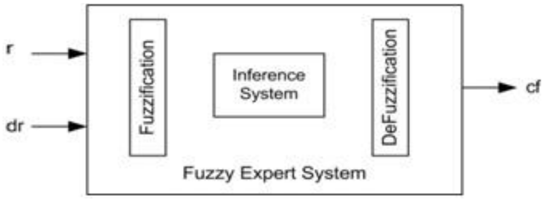


Fig. 5. Schema bloc of FES

In fact, It has two inputs: residue/error (**r**) and residue derivative (**dr**). The residue is considered in this case as distance separating variables actual/real values from desired/optimal ones. Then, we could define second input as residue derivative that could inform about residue evolution. The output, in another hand, is certitude factor (**cf**) that evaluates the state of concerned variable. Matlab is used to develop and to simulate FES because it had fuzzy logic toolbox.

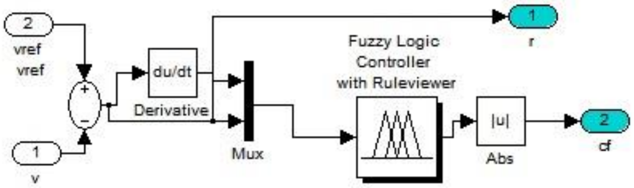


Fig. 6. Matlab Schema bloc of r and dr calculation

A. Fuzzification

Input variables could be qualified as table 2 indicates. Each variable could be presented by linguistic terms (3-5-7). Three scenarios are to be considered:

TABLE II. FUZZIFICATION R AND DR

Prop.	*Number of Membership functions*	*Symbolic labels*	*Types of membership functions*
Sc 1	3 A(r)/B(dr)= {N,Z,P}	N (negatif) Z (zero) P (positif)	Trapezoïdal/ Triangular
Sc 2	5 A(r)/B(dr)= {NB,NS,Z,PS, PB}	NB (negatif big) NS (negatif small) Z (zero) PS (positif small) P (positif big)	Trapezoïdal/ Triangular
Sc 3	7 A(r)/B(dr)= {NB,NM,NS,Z,PS ,PM,PB}	NB (negatif big) NM (negatif moyen) NS (negatif small) Z (zero) PS (positif small) PM (positif moyen) P (positif big)	Trapezoïdal/ Triangular

r and dr are variables ranging respectively in sets of symbolic labels A(r) and B(dr). The terms describe qualitative value of magnitude of both residue and its variations. Fuzzification of the two inputs could adopt three scenarios with: 3 membership functions, 5 membership functions, 7 membership functions.

TABLE III. FUZZIFICATION OF CF

Number of MF	**Linguistic Terms**	**MF types**
6 C(cf)= { AL0, AL0.2, AL0.4, AL0.6, AL0.8, A}	AL0 AL0.2 AL0.4 AL0.6 AL0.8 AL1	triangular

The linguistic variable cf is output variable ranging in sets of symbolic labels C(cf) = {AL0, AL0.2, AL0.4, AL0.6, AL0.8, AL1}.(as table 3).

B. Defuzzification

In this system output calculation, a crisp value is required. Thus, the defuzzification operation is requisite. In this approach, the *gravity centre* is the method adopted to get the crisp value traducing the severity of generated alarm, from the output membership function.

C. Inference engine

Linguistic model relating variables r and dr to variable D is written as rule base, relating the terms of A(r) and B(dr) to those of C(cf) in *n* rules :

If r is A_k and dr is B_l then cf is C_s **(1)**

In our research, we are studying the effect of raising the number of symbolic labels describing the linguistic variable. We suppose these hypotheses:

H1: Raising symbolic labels enhance robustness against uncertainty.

H2: It could lead to the augmentation of rule number which affects negatively time response.

We are working within three scenarios depending on the number of membership functions representing variables. For each case, an inference system relating inputs to outputs is proposed.

- FES1 : r (5 MF) and dr (3MF):

In this case, residual associated with five symbolic labels and dr with three symbolic labels. The number of rules n=15. Table 4 is an illustration of the inference.

TABLE IV. FES1: INFERENCE SYSTEM

r	dr		
	N	Z	P
NB	AL1	AL1	AL0.8
NS	AL1	AL0.6	AL0.4
Z	AL0.2	AL0	AL0.2
PS	AL0.4	AL0.6	AL1
PB	AL0.8	AL1	AL1

- FES2 : r (5 MF) and dr (5MF)

In this case, residual associated with five symbolic labels and dr with five symbolic labels. The number of rules n=25. Table 5 is an illustration of the inference.

TABLE V. FES2: INFERENCE SYSTEM

r	Dr				
	NB	NS	Z	PS	PB
NB	AL1	AL1	AL1	AL0.8	AL0.6
NS	AL1	AL0.6	AL0.2	AL0.6	AL0.4
Z	AL0.4	AL0.2	AL0	AL0.2	AL0.4
PS	AL0.4	AL0.6	AL1	AL0.6	AL1
PB	AL0.6	AL0.8	AL1	AL1	AL1

- FES3: r (7MF) and dr (7MF)

In this case, residual associated with seven symbolic labels and dr with seven symbolic labels. The number of rules n=49. Table 6 is an illustration of the inference.

TABLE VI. FES3: INFERENCE SYSTEM

r	dr						
	NB	NM	NS	Z	PS	PM	PB
NB	AL1	AL1	AL1	AL1	AL0.8	AL0.8	AL0.6
NM	AL1	AL1	AL0.8	AL0.6	AL0.6	AL0.4	AL0.4
NS	AL0.8	AL0.6	AL0.4	AL0.4	AL0.2	AL0.2	AL0
Z	AL0.4	AL0.2	AL0	AL0	AL0	AL0.2	AL0.4
PS	AL0	AL0.2	AL0.2	AL0.4	AL0.4	AL0.6	AL0.8
PM	AL0.4	AL0.4	AL0.6	AL0.6	AL0.8	AL1	AL1
PB	AL0.6	AL0.8	AL0.8	AL1	AL1	AL1	AL1

- Comparative study and results

To compare and define conclusions, we fix the universe of discourse inputs and output: r is in [-2 2], dr is in [-8 8]. We should mention that these intervals depend on studied situations and variables.

Assuming that the construction of three inference system obeyed to the same logic which is: When residual is zero the variable is normal. Otherwise, negative or positive values are synonyms of abnormality. Derivative magnitude informs about residual evolution. Next figures 7, 8 and 9 illustrate 3 d response of three FES.

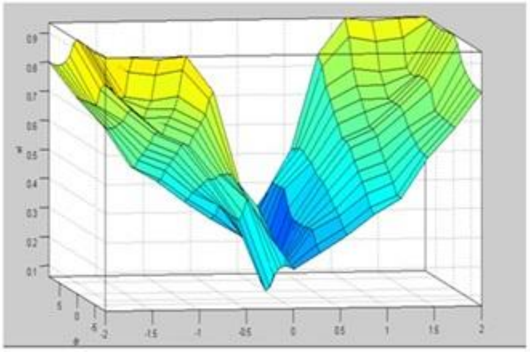


Fig. 7. FES1: 3D cf(r,dr)

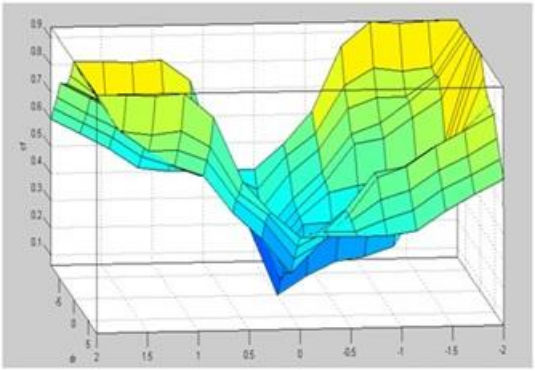


Fig. 8. FES2: 3D cf(r,dr)

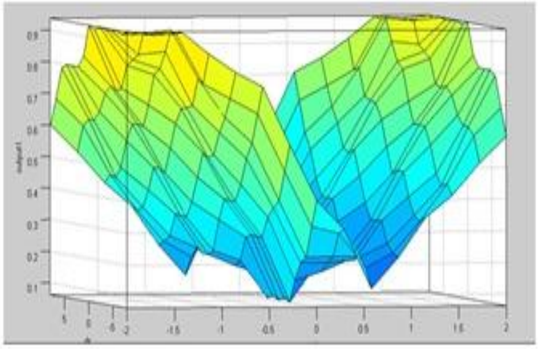


Fig. 9. FES3: 3D cf(r,dr)

We could notice that the three systems have the same evolution: cf is around zero when residual is null and it raised to reach 1 when residual absolute value rises. However, when number of MF is important system response is more slow and soft.

Let's study with precision the three fuzzy expert systems for minimal and for important variations of residual.

For minimal variations of residual, we remark that the three systems don't reach zero even when *cf* is equal to 0. Minimal value is 0.06: this fact is justified by the uncertainty of measures and of information. Around zero the third system is more precise and the confidante zone, where the variable is normal, is larger than the other FES.

TABLE VII. FES1: CF FOR SMALL VARIATIONS OF RESIDUES

r dr	***-0.2***	***-0.1***	***-0.05***	***0***	***0.05***	***0.1***	***0.15***	***0.2***
-8	0,27	0,2	0,2	0,2	0,2	0,2	0,21	0,23
-6	0,33	0,19	0,19	0,19	0,19	0,19	0,24	0,27
-4	0,32	0,17	0,17	0,17	0,17	0,17	0,22	0,26
-2	0,32	0,12	0,12	0,12	0,12	0,12	0,19	0,24
0	0,23	**0,06**	**0,06**	**0,06**	**0,06**	**0,06**	0,16	0,23
2	0,24	0,12	0,12	0,12	0,12	0,12	0,24	0,32
4	0,26	0,17	0,17	0,17	0,17	0,17	0,26	0,32
6	0,27	0,19	0,19	0,19	0,19	0,19	0,27	0,33
8	0,23	0,2	0,2	0,2	0,2	0,2	0,24	0,27

TABLE VIII. FES2: CF FOR SMALL VARIATIONS OF RESIDUES

r dr	***-0.2***	***-0.1***	***-0.05***	***0***	***0.05***	***0.1***	***0.15***	***0.2***
-8	0,45	0,4	0,4	0,4	0,4	0,4	0,4	0,4
-6	0,38	0,3	0,3	0,3	0,3	0,3	0,31	0,33
-4	0,27	0,2	0,2	0,2	0,2	0,2	0,24	0,27
-2	0,26	0,17	0,17	0,17	0,17	0,17	0,26	0,32
0	0,10	**0,06**	**0,06**	**0,06**	**0,06**	**0,06**	0,15	0,22
2	0,26	0,17	0,17	0,17	0,17	0,17	0,26	0,32
4	0,27	0,2	0,2	0,2	0,2	0,2	0,24	0,27
6	0,33	0,3	0,3	0,3	0,3	0,3	0,34	0,38
8	0,4	0,4	0,4	0,4	0,4	0,4	0,42	0,45

TABLE IX. FES3: CF FOR SMALL VARIATIONS OF RESIDUES

r dr	*-0.2*	*-0.1*	*-0.05*	*0*	*0.05*	*0.1*	*0.15*	*0.2*
-8	0,52	0,44	0,4	0,4	0,4	0,38	0,35	0,33
-6	0,39	0,29	0,24	0,24	0,24	0,24	0,24	0,24
-4	0,31	0,22	0,17	0,17	0,17	0,17	0,17	0,17
-2	0,22	0,13	**0,06**	**0,06**	**0,06**	0,13	0,18	0,20
0	0,22	0,12	**0,06**	**0,06**	**0,06**	0,12	0,18	0,22
2	0,20	0,13	**0,06**	**0,06**	**0,06**	0,13	0,18	0,22
4	0,17	0,17	0,17	0,17	0,17	0,22	0,28	0,31
6	0,24	0,24	0,24	0,24	0,24	0,29	0,36	0,39
8	0,33	0,38	0,4	0,4	0,4	0,44	0,49	0,52

In this case, we could conclude on the fact that first and second systems are more suitable for critical situations where little variations of residues are significant for system safety. However, the third system could be best used in non critical situations.

TABLE X. FES1: CF FOR IMPORTANT VARIATIONS OF RESIDUES

r dr	***-2***	***-1.5***	***-1***	***0***	***0.5***	***1***	***1.5***	***2***
-8	**0,93**	0,92	0,93	0,2	0,4	0,4	0,5	**0,6**
-6	**0,92**	0,71	0,71	0,19	0,4	0,5	0,6	**0,7**
-4	**0,93**	0,71	0,6	0,17	0,4	0,6	0,7	**0,8**
-2	**0,92**	0,50	0,4	0,12	0,47	0,71	0,72	**0,82**
0	**0,93**	0,44	0,2	0,06	0,5	0,93	0,92	**0,93**
2	**0,82**	0,54	0,4	0,12	0,47	0,71	0,71	**0,92**
4	**0,8**	0,7	0,6	0,17	0,4	0,6	0,71	**0,93**
6	**0,7**	0,6	0,5	0,19	0,49	0,71	0,71	**0,92**
8	**0,6**	0,5	0,4	0,2	0,57	0,93	0,92	**0,93**

TABLE XI. FES2: CF FOR IMPORTANT VARIATIONS OF RESIDUES

r dr	***-2***	***-1.5***	***-1***	***0***	***0.5***	***1***	***1.5***	***2***
-8	**0,93**	0,92	0,93	0,2	0,3	0,4	0,6	**0,8**
-6	**0,93**	0,77	0,79	0,19	0,34	0,44	0,60	**0,80**
-4	**0,92**	0,71	0,71	0,17	0,37	0,5	0,62	**0,82**
-2	**0,93**	0,71	0,64	0,12	0,40	0,55	0,65	**0,87**
0	**0,93**	0,71	0,6	0,06	0,42	0,6	0,71	**0,93**
2	**0,87**	0,65	0,55	0,12	0,46	0,64	0,71	**0,93**
4	**0,82**	0,62	0,5	0,17	0,47	0,71	0,71	**0,92**
6	**0,80**	0,60	0,44	0,19	0,47	0,79	0,77	**0,93**
8	**0,8**	0,6	0,4	0,2	0,44	0,93	0,92	**0,93**

TABLE XII. FES3: CF FOR IMPORTANT VARIATIONS OF RESIDUES

r dr	***-2***	***-1.5***	***-1***	***0***	***0.5***	***1***	***1.5***	***2***
-8	**0.94**	**0.93**	0.83	0.70	0.40	0.20	0.45	**0.6**
-6	**0.93**	**0.93**	0.71	0.54	0.25	0.24	0.51	**0.75**
-4	**0.92**	0.83	0.63	0.44	0.17	0.19	0.55	**0.80**
-2	**0.93**	0.76	0.60	0.35	0.07	0.24	0.65	**0.80**
0	**0.94**	0.65	0.50	0.35	0.06	0.35	0.65	**0.94**
2	**0.80**	0.65	0.40	0.24	0.07	0.35	0.76	**0.93**
4	**0.80**	0.55	0.40	0.19	0.17	0.44	0.83	**0.92**
6	**0.75**	0.51	0.29	0.24	0.25	0.54	**0.93**	**0.93**
8	**0.60**	0.45	0.29	0.20	0.40	0.70	**0.93**	**0.94**

For important variation of residual, three systems are reaching their maximum value which is 0.93. This coefficient is synonym of evident abnormal situation for the considered variable.

To highlight these results, FES proposed must be applied on systems from different fields. The next section is reserved to three applications of FES for surveillance: industrial and medical one.

V. STUDY CASE

The aim of this work is to propose generic fuzzy expert system that could monitor several kinds of situations.

A. *Industrial application : Three tanks system*

The system under consideration is a pilot plant of the research unit: *System analysis and command* located in ENIT (National Engineer Institute of Tunisia). The considered system is composed of three interconnected cylindrical tanks, two pumps, six valves, pipes, water reservoir in the bottom, measurement of liquid levels and other elements. The pumps pump water from the bottom reservoir to the top of the left and right tanks.

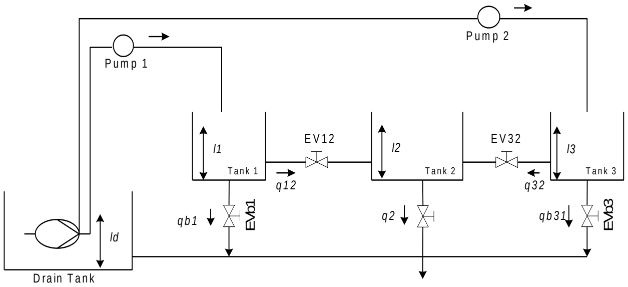


Fig. 10. Three tanks model

- <u>System Modeling</u>

While tanks 1, 2 and 3 are identical with cross section S and maximum fluid level l_{max}, Drain tank is characterized with cross section S_d and maximum fluid level l_{dmax}. Tanks 1 and 3 are coupled with tank 2 by two AON (all-or-none) valves with cross section S_n and outflow coefficients. Two proportional valves EV1 and EV2 directly connected to a pump, with highest possible flow rate denoted q_{max} supply tanks 1 and 2. Three sensors are installed to measure the three levels l_1, l_2 and l_3. The experimental plant that is equipped with sensors and actuators, communicates via data acquisition system with a personal computer.

Because the modeled system presents much non linearity, we've tried to model each component in separate block. All individual parts models were incorporated into single block in Matlab/Simulink environment. The block has 11 inputs: 2 float signals controlling the pumps (1 and 2) and 9 Boolean signals controlling the valves (EV12, EV32 and EV2). Besides it has 3 float signals outputs from water level heights.[58]

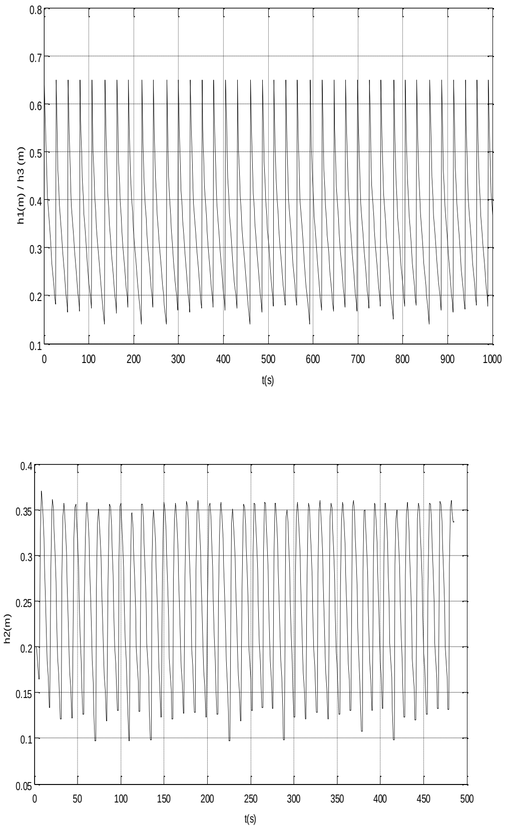


Fig. 11. L1/l2/l3 optimal variations

- FES application and results:

Measurable variables that could inform about system state are summarized in table 13. After fixing studied variables, residues and their variations are calculated. In this system we are installing three FES associated with all variables. In this section, we suppose that studied system is having an actuator failure in the tank 1 that leaks. Tank1 leaking is happening at the 70th second and its value is equal to *0.02*. We are studying in the following paragraphs three FES responses to simulated failure. Figure 12 is an illustration of certitude coefficient calculated for tank1 level by both 3 proposed FES.

TABLE XIII. MEASURABLE VARIABLES

Variables	*Name*	*Normal values*	*Range residues*	*Associate anomalies*
Tank1 level	*l1*	*h1 optimal variations*	*[-1 1]*	*f1: Tank1 leak* *f2: EV12 failed* *f3:Pump1 failed*
Tank2 level	*l2*	*l2 optimal variations*	*[-1 1]*	*f4:Tank2 leak* *f5:EV12 failed* *f6:EV23 failed*
Tank3 level	*l3*	*l3 optimal variations*	*[-1 1]*	*f7:Tank3 leak* *f8:EV23 failed* *f9:Pump3 failed*

It is remarkable that they respond with the same shape and variations while they obey to the same logic. Alarm is generated immediately when cf is superior to 0.2. Figure 13 illustrates that: FES1 generates alarm at 70.27 s time, FES2 at 70.275 s and FES3 at 70.29 s. However, FES1 is having most important maximum value (cf reaches 0.625) and FES 2 and FES 3 reaches 0.6 for maximum cf value. These results mean that an alarm is generated in appropriate time for variable l1 that is immediately related with leak.

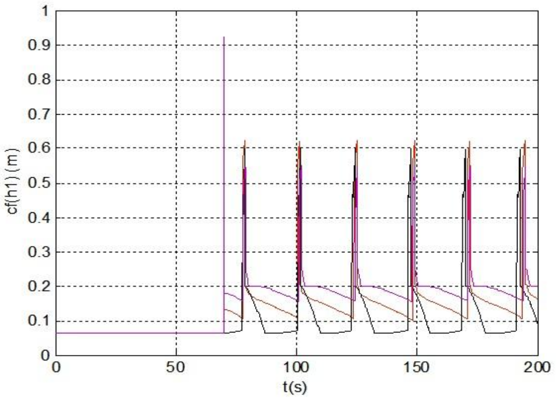


Fig. 12. Cf(t) (-:FES1, -: FES2, -:FES3)

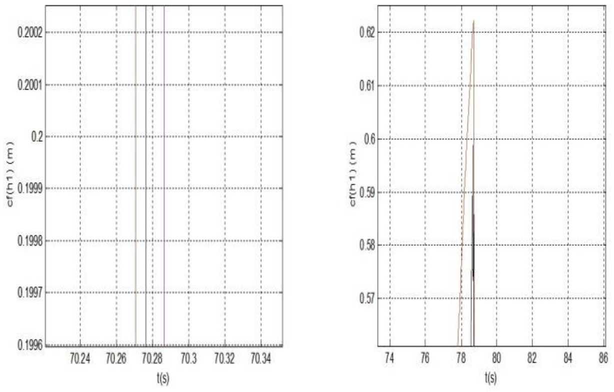


Fig. 13. Cf(t) maximum and minimum values(-:FES1, -: FES2, -:FES3)

Looking at first alarm generation time and alarm maximum severity factor, fault localization is evident.

Variable l1(t) that monitors Tank 1 is the most and the first affected one. Consequently, this decision support system guides human operator in his decision. To diagnosis the current situation, a normalized fault signature depending on calculated certitude coefficient is proposed. Having a vector r(k) (k in {1,2,3}) describing respectively residues of the three tanks levels, we define a normalized vector $r_n(k)$:

$$r_n(k) \ = \left\{ \begin{matrix} 0 & if \ \ cf(r(k)) < 0.2 \\ 1 & if \ \ cf(r(k)) > 0.2 \end{matrix} \right\}$$

A failure signature matrix could indicate about the incidence of failures on residues:

$$\begin{bmatrix} r1n \\ r2n \\ r3n \end{bmatrix} \cong (depend\ on) \begin{bmatrix} 1\,1\,1\,0\,1\,0\,0\,0\,0\,1 \\ 0\,1\,0\,1\,0\,1\,0\,0\,1\,0 \\ 0\,0\,0\,1\,0\,0\,1\,1\,1\,0 \end{bmatrix} \begin{bmatrix} f1 \\ f2 \\ f3 \\ f4 \\ f5 \\ f6 \\ f7 \\ f8 \\ f9 \\ \end{bmatrix}$$

In this study case, we conclude that the generic fuzz expert system applied to industrial field provides a decision support system. It detects failures and generates alarms with severity or certitude factors in one hand. In another hand, it helps locating failure origin root based on certitude factor and first alarm generation time. Industrial field supposes that *n* (number of variables characterizing system) FES has to be installed. These FES are running in real time while the studied process is functional which makes the first FES the more suitable to be applied because of its time response and time execution.

TABLE XIV. MEASURABLE VARIABLES

	First alarm generation time	**Alarm max certitude/s everiy**	**Alarm persistence mean time**	**Fault localization**
l1	**FES1** :70.27s **FES2** :70.275s **FES3**:70.29s	62% 60% 60%	1.04 s	Failure root
l2	**FES1** :75s **FES2** :76s **FES3**:77s	26% 19% 23%	0.06 s	Failure propagation
l3	**FES1** :101s **FES2** :102s **FES3**:104s	21% 22% 22%	0.052s	Failure propagation

B. *Medical application : ECG analysis*

An electrocardiogram (ECG) is a simple and commonly performed test that records the electrical activity of the heart. An ECG is used to measure the rate and rhythm of the heart. It is a useful investigation in screening for heart disease and for those people who have a cardiovascular disorder. An ECG can show the presence of any damage to the heart, although not all heart conditions can be detected by an ECG.

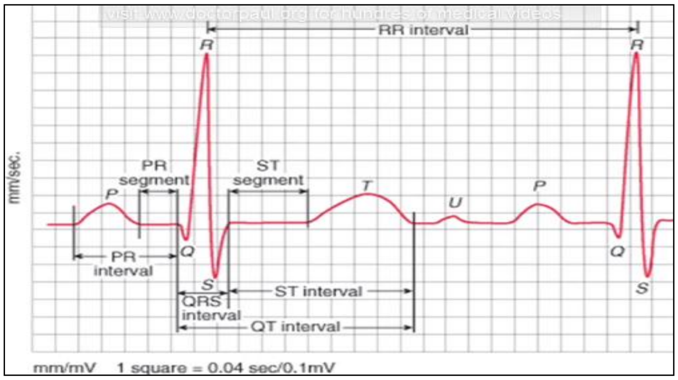


Fig. 14. ECG signal

Table 14 summarizes the inputs of this module which are deduced from the signal in figure 14 of ECG. The table gives also normal values of the inputs. Limits and thersholds for normal values are those of an athlete [61].

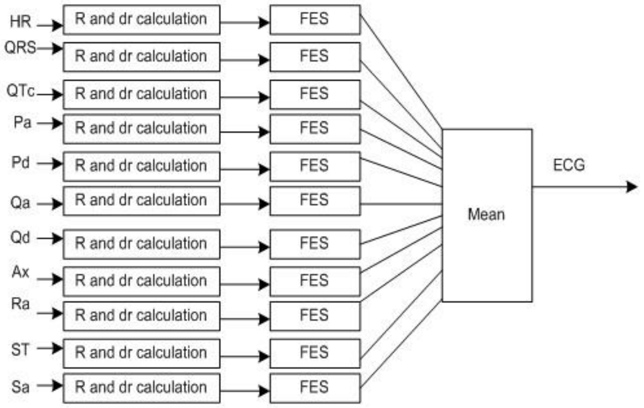


Fig. 15. ECG monitoring system

The calculation of residue is as the folowing equations shows :

If $v \epsilon [l_{min} - l_{max}]$ *then* $r = 0$
If $v < l_{min}$ *or then* $r = v\text{-}\ l_{min}$
$v > l_{max}$ *then* $r = v\text{-}l_{max}$

Considering that t_i is the actual date of analysis, $t_{i-1}$is the last one, so we could calculate:

$$dr_{\ i} = [r_i - \ r_{i-1}]/t_i \ - t_{i-1}$$

TABLE XV. VARIABLES CHARECHTERIZING ECG

Variables	*Name*	*Normal Values for an athlete*	*Range of r and dr*	*AssociatedAn omalies ofl ECG*
Heart rate	HR	[45 – 150]	r [-4 4] dr [-2 2]	Bradycadie Tachycardie
QRS	QRS	QRS<0.12s	r [-0.1 0.1] dr [-2 2]	Complete bundle brunch block
QTc	QTc	0.34>QTc>=0.48s	r [-0.1 0.1] dr [-2 2]	Short QTc interval Long QTc interval
P wave amplitude	Pa	Pa<2.5mm	r [-1 1] dr [-2 2]	Right atrial enlargement
P wave duration	Pd	Pd<0.04s	r [-0.1 0.1] dr [-2 2]	left atrial enlargement
Q wave amplitude	Qa	Qa<3 mm	r [-1 1] dr [-2 2]	Pathologic Q wave
Q wave duration	Qd	Qd<0.04s	r [-0.1 0.1] dr [-2 2]	Pathologic Q wave
Axis	Ax	-90°<Ax<120°	r [-4 4] dr [-2 2]	-Left axis deviation -Right ventricular hypertrophy
R wave amplitude	Ra	Ra<5 mm	r [-1 1] dr [-2 2]	Pathologic R wave
ST segment	ST	ST<1 mm	r [-1 1] dr [-2 2]	ST segment depression
T wave amplitude	Sa	Sa<1 mm	r [-1 1] dr [-2 2]	T wave inversion

- FES application and results

While ECG is normal, the three fuzzy expert systems response is illustrated in table 15.

TABLE XVI. RESULTS WITH NORMAL ECG

Variables	Residue	Residues derivative	cf SEF 1	cf SEF 2	cf SEF 3
HR=110	0	0	0.063	0.063	0.063
QRS=0.1	0	0	0.063	0.063	0.063
QTc=0.4	0	0	0.063	0.063	0.063
Pa=2	0	0	0.063	0.063	0.063
Pd=0.02	0	0	0.063	0.063	0.063
Qa=2	0	0	0.063	0.063	0.063
Qd=0.035	0	0	0.063	0.063	0.063
Axis=35	0	0	0.063	0.063	0.063
Ra=3	0	0	0.063	0.063	0.063
ST=0.51	0	0	0.063	0.063	0.063
Sa= 0.5	0	0	0.063	0.063	0.063
ECG Normal - cf			94 %	94 %	94 %
Time response			0.0180	0.021	0.029
Precision			0.063	0.063	0.063

TABLE XVII. RESULTS WITH ABNORMAL ECG

Variables	cf SEF 1	cf SEF 2	cf SEF 3
HR=160	0.063	0.063	0.063
QRS= 0.18	0.600	0.611	0.685
QTc=0.483	0.2758	0.334	0.319
Pa=2	0.063	0.063	0.063
Pd=0.02	0.063	0.063	0.063
Qa=2	0.063	0.063	0.063
Qd=0.035	0.063	0.063	0.063
Axis=35	0.063	0.063	0.063
Ra=3	0.063	0.063	0.063
ST=0.51	0.063	0.063	0.063
Sa= 0.5	0.063	0.063	0.063
ECG Normal - cf	40 %	39 %	32 %
Time response	0.0190	0.018	0.025
ECG certitude coefficient	0.600	0.611	0.685

The whole system output is in this case the maximum between eleven calculated cf. If system output is superior then 0.2, ECG is abnormal. In the same way, ECG diagnosis could be done using heart anomalies signatures. In this case signature is equal to [0 1 1 0 0 0 0 0 0 0 0] which is equivalent to “Complete bundle brunch block”

VI. CONCLUSION

This research work aims to study fuzzy expert system for monitoring. These support decision systems are very used in several applications and fields. A state of art has proved that developed fuzzy expert systems are usually specific to studied process whether in variables choice or in inference logic.

Proposing generic fuzzy expert system for surveillance independently of its application had been the subject of this paper. The main idea is to characterize concerned application with measurable variables. Fuzz expert system is installed to monitor each one by analyzing the distance between variable and its optimum behavior (error or residue).

Three generic fuzzy expert systems were proposed. The number of membership functions describing error and its variations differentiates between different proposed systems. Two criteria had been discussed time response and incertitude minimization. Increasing membership functions improves precision and describes better each variable variation. However, system time response rises which is annoying especially in real time.

When applying FES in industrial and medical diagnosis, results confirms that it provides decision support systems that detects abnormal situations and affects certitude coefficient enhancing uncertainty.

VII. FUTURE WORKS

Diagnosis process which is in our approach based on anomalies signatures could be improved by using artificial neural networks (ANN). Multi layer perceptron (MLP) is a suitable tool for anomalies classification that had been used in many applications.

REFERENCES

[1] Dubois D., Prade H., “An Introduction to fuzzy systems,” Clinica Chimica Acta, Vol. 270, pp. 3-29, 1998.

[2] Bellman R. E., Zadeh L., “Decision making in a fuzzy environment,” Manage Sc, Vol. 17, pp. 141-146, 1970.

[3] Dubois D., Fargier H., Prade H.“ Refinements of the maximum approach to decision making in a fuzzy environment,” Fuzzy Sets and Systems, Vol. 81, pp. 3-29, 1996.

[4] Sahin S., Tolun M.R., Hassanpour R., “Hybrid expert systems: A survey of current approaches and application,” Expert Systems with Applications, vol. 39, p. 4609–4617, 2012.

[5] Siler W., Buckley J., “Fuzzy Expert Systems and Fuzzy Reasoning,” John Wiley & Sons, Inc., New Jersey, 2005.

[6] Venkatasubramanian V., Rengaswamy R., S.N. Kavur, Y. Kewen, “A review of process fault detection and diagnosis Part II: Qualitative Models and search strategie,” Computers and Chemical Engineering, Vol. 27, pp. 313-326, 2003.

[7] A. Evsukoff, S., Gentil, J., Montmain, “Fuzzy reasoning in co-operative supervision systems,” Control Engineering Practice, Vol. 8, pp. 389-407, 2000.

[8] Németh B., Laboncz S., Kiss I., Csépes G, “Transformer condition analyzing expert system using fuzzy neural system,” IEEE International Symposium on Electrical Insulation (ISEI), Canada, 2010.

[9] Yun-Seong L., Hyung-Chul K., Jun-Min C., Jin-O, “A New Method for FMECA Using Expert System And Fuzzy Theory,” PMAPS, 2010.

[10] H.C.W. Lau, R.A. Dwight, “A fuzzy-based decision support model for engineering asset condition monitoring – A case study of examination of water pipelines,” Elsevier -Expert Systems with Applications, vol. 38, pp. 13342–13350, 2011.

[11] Abdelazeem A., Abdelsalam A., Eldesouky Abdelhay, A., Sallam, “ Characterization of power quality disturbances using hybrid technique of linear Kalman filter and fuzzy-expert system,” Elsevier -Electric Power Systems Research, vol. 83, pp. 41– 50, 2012.

[12] Mohsen Naderpour, Jie Lu, “A Fuzzy Dual Expert System for Managing Situation Awareness in a Safety Supervisory System,” WCCI 2012 IEEE World Congress on Computational Intelligence, pp. 10-15, 2012.

[13] Petr Chalupa, Jakub NOVÁKovak, Vlademir Bonbal, “ Detailed Simulink Model of Real Time Three Tank System,” Recent Researches in Circuits, Systems, Communications and Computers, pp. 161-166, 2010.

[14] Abdelaziz, A.Y., Mekhamer, S.F., Nada, M.H., “A fuzzy expert system for loss reduction and voltage control in radial distribution systems,” Elsevier - Electric Power Systems Research, vol. 80, p. 893–897, 2010.

[15] Angulo C., Cabestany J., Rodríguez P., Batlle M., González A., de Campos S., “Fuzzy expert system for the detection of episodes of poor water quality through continuous measurement,” Elsevier -Expert Systems with Applications, vol. 39, pp. 1011–1020, 2012.

[16] Dan-yang C., Shui-ping Z., Jin-hong L., “Variable universe fuzzy expert system for aluminum electrolysis,” Elsevier -Trans. Nonferrous Met. Soc, vol. 21, p. 429-436, 2011.

[17] Krivoulya G., Dudar Z., Kucherenko D., Mehana S., “Fuzzy Expert System for Diagnosis of Computer Failures,” CADSM’2009, 24-28 February, Polyana-Svalyava, UKRAINE, p. 225-230, 2009.

[18] Chang-Shing Lee, Mei-Hui Wang, “A Fuzzy Expert System for Diabetes Decision Support Application,” IEEE Transaction on Systems Man and Cybernetics - PART B: Cybernetic, Vol. 41, N°. 1, pp. 139-153, 2011.

[19] Azian Azamimi Abdullah, Zulkarnay Zakaria and Nur Farahiyah Mohammad, “Design and Development of Fuzzy Expert System for Diagnosis of Hypertension,” Second International Conference on Intelligent Systems, Modelling and Simulation (ICISMS 2011), pp. 113-117, 2011.

[20] M. Neshat, M. Yaghobi, M. B. Naghibi, A. Esmaelzadeh, “Fuzzy Expert System Design for Diagnosis of liver disorders,” International Symposium on Knowledge Acquisition and Modeling (ISKAM 2008), pp. 252-256, 2008.

[21] M.J.P. Castanho, F. Hernandes, A.M. De Ré, S. Rautenberg, A. Billis, “Fuzzy expert system for predicting pathological stage of prostate cancer,” Elsevier -Expert Systems with Applications, 2012.

[22] Ali Adeli & Adeli Neshat, “A Fuzzy Expert System for Heart Disease Diagnosis,” Proceedings of the international multi conference of engineering and computer scientists, hong kong, mars 17-19 -2010.

[23] Ali Keles, Aytürk Keles, Ugur Yavuz, “Expert system based on neuro-fuzzy rules for diagnosis breast cancer,” Expert Systems with Applications, vol. 38, pp. 5719–5726, 2011.

[24] I.M. Dokas, D.A. Karras, D.C. Panagiotakopoulos, “Fault tree analysis and fuzzy expert systems: Early warning and emergency response of landfill operations,” Environmental Modelling & Software, vol. 24, pp. 8–25, 2009.

[25] M.H. Fazel Zarandi, P. Ahmadpou, “Fuzzy agent-based expert system for steel making process,” Expert Systems with Applications, vol. 36, pp. 9539–9547, 2009.

[26] Chou S., Chang Y., Shen, C., “A fuzzy simple additive weighting system under group decision-making for facility location selection with objective/subjective attributes,” European Journal of Operational Research, vol. 198, pp. 132-145, 2008.

[27] Dell’Orco M., Circella G., Sassanelli D., “A hybrid approach to combine fuzziness and randomness interval choice prediction,” European Journal of Operational Research, vol. 195, pp. 648-658, 2008.

[28] Doukas H. C., Andreas B. M. Psarras J. E., “Multi-criteria decision aid for the formulation of sustainable technological energy priorities using linguistic variables,” European Journal of Operational Research, vol. 182, pp. 844-855, 2007.

[29] Deyin Ma., Yanchun, L. Xiaoshe, Z. Renchu G. Xiaohu S, “Multi-BP expert system for fault diagnosis of power system,” Engineering Applications of Artificial Intelligence, vol. 26, pp. 937–944, 2013.

[30] O.C. Pires, C. Palma, I. Moita, J.C. Costa, M.M. Alves and E.C. Ferreira, “A fuzzy logic based expert system for diagnosis and control of an integrated wastewater treatment,” 4th Mercosur Congress on Process Systems Engineering, Brasil, 2013.

[31] John Farrell, Abraham Kandel, “A fuzzy rule-based expert system for marine bioassessment,” Fuzzy Sets and Systems, vol. 89, pp. 27-34, 1997.

[32] G. Hong, X. Chen, X. Xue, S. Zhang, “Expert Systems for Fault Diagnosis Integrating Neural Network and Fuzzy Inference,” International Conference of Information Technology, Computer Engineering and Management Sciences, pp. 245-249, 2011.

[33] G. Molnárka, “Management of Uncertainty in Visual Examination Procedure in Building Diagnostics with Fuzzy Expert System,” ISCIII 2009: 4th International Symposium on Computational Intelligence and Intelligent Informatics, pp. 21–25, Egypt, 2009.

[34] LI Jie, SHEN Shi-tuan, “Research on the Algorithm of Avionic Device Fault Diagnosis Based on Fuzzy Expert System,” Chinese Journal of Aeronautics, Vol. 20, pp. 223-229, 2007.

[35] Liu Xiaobo, Li Jianping, “Fault Diagnosis of Fan Based on Fuzzy Neural Expert System,” Second International Conference on Intelligent Computation Technology and Automation, 2009.

[36] Chi-Jen Lin a, Wei-Wen Wu, “A causal analytical method for group decision-making under fuzzy environment,” Expert Systems with Applications, vol. 34, pp. 205–213, 2008.

[37] Jacky Montmain, Sylviane Gentil, “Dynamic causal model diagnostic reasoning for online technical process supervision,” Automatica, vol. 36, pp. 1137-1152, 2000.

[38] Bodapati Nageswararao & B. Jeyasurya, “Fuzzy-expert system for voltage stability monitoring and control,” Electric Power Systems Research vol. 47 pp. 215–222, 1998.

[39] Burak Ozyurt, Abraham Kandel, “A hybrid hierarchical neural network-fuzzy expert system approach to chemical process fault diagnosis,” Fuzzy Sets and Systems, vol. 83 pp. 11- 25, 1996.

[40] P. Baraldi, M. Librizzi, E. Zio, L. Podofillini, V.N. Dang, “Two techniques of sensitivity and uncertainty analysis of fuzzy expert

systems," Expert Systems with Applications, vol. 36, pp. 12461–12471, 2009.

[41] Palluat N., Racoceanu D., Zerhouni N., "A neuro-fuzzy monitoring systemApplication to flexible production systems," Computers in Industry, vol. 57, pp. 528–538, 2006.

[42] Yvonne Power & Parisa A. Bahri, "A two-step supervisory fault diagnosis framework," Computers and Chemical Engineering, vol. 28, pp. 2131–2140, 2004.

[43] Ling Wang, Jian Chu, Jun Wu, "Selection of optimum maintenance strategies based on a fuzzy analytic hierarchy process," Int. J. Production Economics, vol. 107, pp. 151–163, 2007.

[44] Zhang Quan, Chen Nanyu, Huang Jun, Meng Zhijun, "Application of Expert System Fuzzy BP Neural Network in Fault Diagnosis of Piston Engine," International Conference on Computer Science and Electronics Engineering, pp.604-607, 2012.

[45] M. Kalpana, A.V Senthil Kumar, "Fuzzy Expert System for Diabetes using Fuzzy Verdict Mechanism," Int. J. Advanced Networking and Applications, Vol. 03(02), pp. 1128-1134, 2011.

[46] Mahdi Jampour, Mohsen Jampour, Maryam Ashourzadeh, Mahdi Yaghoobi, "A Fuzzy Expert System to Diagnose Diseases with Neurological Signs in Domestic Animal," 2011 Eighth International Conference on Information Technology: New Generations, 2011.

[47] Novruz ALLAHVERDI & Tevfik AKCAN, "A Fuzzy Expert System Design for Diagnosis of Periodontal Dental Disease," 5 th international conference on application of information and communication technologies, 2011.

[48] Oana GEMAN, "A Fuzzy Expert Systems Design for Diagnosis of Parkinson's Disease," Proceedings of the 3rd International Conference on E-Health and Bioengineering - EHB 2011.

[49] Djam, X.Y. and Y.H. Kimbi, "Fuzzy Expert System for the Management of Hypertension," Pacific Journal of Science and Technology, Vol. 12(1), pp. 390-402, 2011.

[50] Miguel L. J., Blazquez L. F., "Fuzzy Logic based decision making for fault diagnosis in a DC motor," Engineering applications of Artificial Intelligence, vol. 18, pp. 432-450, 2005.

[51] Baron L., Achche S. and Balazinski M., "Fuzzy decision support system knowledge base generation using genetic algorithm," International Journal of approximative reasoning, vol. 28, pp. 125-148, 2011.

[52] M. J. P. Castanho, L. C. de Barros, Akebo Yamakami, Lae´rcio Luis Vendite, "Fuzzy expert system: An example in prostate cancer," Applied Mathematics and Computation, vol. 202, pp. 78–85, 2008.

[53] Azian Azamimi Abdullah, Zulkarnay Zakaria and Nur Farahiyah Mohammad, "Design and Development of Fuzzy Expert System for Diagnosis of Hypertension," Second International Conference on Intelligent Systems, Modelling and Simulation (ICISMS 2011), pp. 113-122217, 2011.

[54] Péter L. Venetianer, Hongli Deng, "Performance evaluation of an intelligent video surveillance system – A case study," Computer Vision and Image Understanding, vol. 114, pp. 1292–1302, 2010.

[55] Beuthe M., Eeckhoudt L., Scannella G., "A practical multicriteria methodology for assessing risky public investments," Socio-Economic Planning Sciences, vol. 34, pp. 121-139, 2000.

[56] Kunch, P. L., Fortemps, P., "A fuzzy decision support system for the economic calculus in radioactive waste management," Information science, vol. 142, pp. 103-116, 2002.

[57] Wang J., & Lin Y., "A fuzzy multicriteria group decision making approach to select configuration items for software development," Fuzzy sets and systems, vol. 134, pp. 343-363, 2003.

[58] Najar Yosra, Ketata Raouf, Ksouri Mekki, "Fuzzy Expert System for Residual Analysis," Journal of Information Organization, Volume 3 Number 1, pp. 23-35, 2013.

[59] Anup Kumar Panda, Suresh Mikkili. "FLC based shunt active filter (p–q and Id–Iq) control strategies for mitigation of harmonics with different fuzzy MFs using MATLAB and real-time digital simulator," Electrical Power and Energy Systems, vol. 47, pp. 313–336, 2013.

[60] Alexandros Iosifidis, Anastasios Tefas, Nikolaos Nikolaidia, Ioannis Pitas, "Multi-view human movement recognition based on fuzzy distances and linear discriminant analysis," Computer Vision and Image Understanding, vol. 116, pp. 347–360, 2012.

[61] Jonathan A Drezner, "Standardised criteria for ECG interpretation in athletes: a practical tool," Br J Sports Med, Vol 46, 2012.

The Determination of Affirmative and Negative Intentions for Indirect Speech Acts by a Recommendation Tree

Takuki Ogawa, Kazuhiro Morita, Masao Fuketa, Jun-ichi Aoe
Dept. of Information Science and Intelligent Systems
University of Tokushima
Tokushima City, Japan

***Abstract*—For context-based recommendation systems, it is necessary to detect affirmative and negative intentions from answers. However, traditional studies cannot determine these intentions from indirect speech acts.**

In order to determine these intentions from indirect speech acts, this paper defines a recommendation tree and proposes an algorithm of deriving intentions of indirect speech acts by the tree. In the proposed method, a recommendation condition (RC) is introduced and it is classified into a required RC, a selectable RC, and a not-selectable RC. The recommendation tree is constructed by nodes and edges corresponding to these three conditions. The deriving algorithm determines affirmative and negative intentions of indirect speech acts by tracing the trees.

From experimental results, it is verified that the accuracy of the proposed method is about 40 points higher than the traditional method.

Keywords—recommendation system; indirect speech acts; affirmative intention; negative intention

I. Introduction

As recommendation systems [1] assist users to select commodities, services, and information, it is necessary to elicit users' requirements in conversational contexts. There are many studies of context-based recommendation systems [2-9]. In these systems with proactive recommendations, affirmative and negative intentions of answers are important in order to decide items [5,6,8,9]. For example, items are commodities in e-commerce sites.

In answers with affirmative and negative intentions, there are direct speech acts and indirect speech acts. The direct speech acts represent these intentions by the following two approaches for a recommendation "How about having a cake today?": the first is fixed phrases such as "O.K." and "No, thank you." without "cakes". The second is sentences representing acceptance and rejection intentions such as "I like cakes." and "I don't want to have cakes." with "cakes", respectively.

In the indirect speech acts, there are two patterns for the recommendation: the first is the affirmative answers that select other cakes excluding chocolate cakes such as "I don't want to have chocolate cakes.". The second is the negative answers that select other foods excluding cakes such as "I want to have Japanese noodles".

In order to determine intentions from sentences, there are two methods: the first uses machine learning such as SVM or HMM, the second uses meaning of words and grammars.

For the machine learning methods which classify sentences into intentions or tag-sets with affirmative and negative ones, Fernandez and Picard [10] divided sentences of Spanish CallHome database (telephone conversations) into eight kinds of dialog act tags [11] by SVM. Surendran and Levow [12] classified sentences of HCRC MapTask corpus [13] into twelve kinds of dialog act tags [14] by SVM and HMM. Stolcke et al. [15] proposed methods of a domain-independent framework for tagging the Discourse Annotation and Markup System of Labeling (DAMSL) tag-set [16] to sentences of conversational speeches. Ravi and Kim [17] classified sentences on discussion boards into six speech act categories by N-gram features and linear SVM. Mera, Ichimura, and Yamashita [18] recognized affirmative and negative intentions from answers of questions by the fuzzy theory.

These methods have the advantage that classification models are constructed automatically, but they expend considerable efforts to collect a large learning data.

For the second classifications by meanings of words and grammars, Kitamura, Watanabe, Sekiguchi, and Suzuki [19] estimated negative intentions by combinations of the following five grammatical features: words, auxiliaries, verbs, adjectives, and superordinate concepts of words in previous sentences. Mera [20] and Yoshie et al. [21] calculated affirmative values of sentences by combining these values of words and formulas of grammars (including modality). These values indicate the strength of affirmative intentions and are defined by questionnaires within [0.0-1.0] scales. For example, affirmative values, "Yes" and "No", are defined as 0.94 and 0.06, respectively. Affirmative formulas reflect effects of modalities in affirmative values. The examples of modalities are adverbs such as "very" and "a little", negative modalities, and double negative modalities.

These methods have the merit that their rules have broad utilities for sentences of many domains, but they can not classify answers of indirect speech acts.

In order to determine affirmative and negative intentions from indirect speech acts, this paper defines a new recommendation tree and proposes a new algorithm of deriving intentions of indirect speech acts by the tree. The tree has a root node which indicates a recommendation. The root node has child nodes corresponding to the following three kinds of recommendation conditions (RC):

1) R_RC is the required RC. In a recommendation, "How about having a cake today?", there are R_RCs, "you", "have", "cake", and "today". In RC "cake", there are R_RCs "sweet" and "sweets".

2) S_RC is the selectable RC. In RC "cake", there are S_RCs of kinds of cakes such as "chocolate cake", "short cake", and "Mont Blanc".

3) NS_RC is the non-selectable RC such as "tomorrow", "Japanese noodle" for the recommendation.

The deriving algorithm determines the intention to the root node from the intention to RCs by tracing the trees. From the indirect speech act "I don't like chocolate cakes." for the above recommendation, the algorithm derives the affirmative intention of the root node from the rejection intention of the S_RC "chocolate cake".

Sections 2 and 3 propose the recommendation tree and the algorithm of deriving intentions, respectively. Section 4 evaluates the proposed method by three kinds of open tests. Section 5 concludes the proposed method.

II. A RECOMMENDATION TREE

In this paper, recommendations include four necessary concepts of RC: "WHO", "WHEN", "WHAT", and "VERB". These concepts have RCs related to persons, schedules, objects, and actions, respectively. Table 1 shows examples of RCs of these concepts.

TABLE I. EXAMPLES OF RCS OF FOUR CONCEPTS

Concepts	RCs
WHO	you he
WHEN	today tomorrow
WHAT	cake curry
VERB	go have

Fig. 1 shows a part of the recommendation tree by using RCs in Table 1. In figures of this paper, node labeled by *x* corresponding to string *x*.

In Fig. 1, root node "REC" indicates the recommendation. The root node has four child nodes corresponding to concepts (concept nodes): "WHO", "WHEN", "WHAT", and "VERB". These nodes have child nodes of RCs (RC nodes). For example, concept node "WHO" has RC nodes "you" and "he". There are three kinds of edges (R_edges, S_edges, and NS_edges) for R_RC, S_RC, and NS_RC, respectively. Root node and concept nodes are connected by R_edges as shown by double lines. Concept nodes and RC nodes are connected by S_edges as shown by single lines. These kinds of edges are changed by each recommendation. For the recommendation "How about having a cake today?", the recommendation tree in Fig. 1 is modified (Fig. 2). In Fig. 2, edges of nodes "you", "today", "cake", and "have" are modified to R_edges. Edges of nodes "he", "tomorrow", "curry", and "go" are set to NS_edges as shown by dotted lines.

The recommendation tree can be extended by expanding terminal nodes. Considering an example in Fig. 3, RC node "cake" in Fig. 1 constructs the subtree as the root node. RC node "cake" has RC nodes "taste" and "kind" with R_edge as child nodes. RC node "taste" has RC node "sweet" with R_edge as a child node. RC node "kind" has RC nodes "Mont Blanc", "short cake", and "chocolate cake" with S_edge as child nodes.

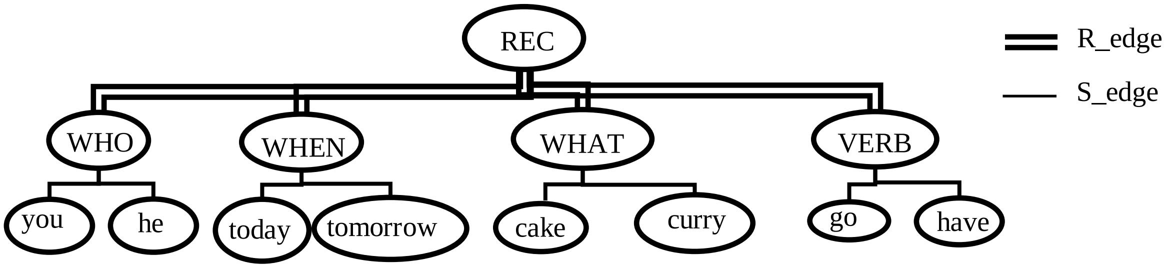


Fig. 1. A part of the recommendation tree

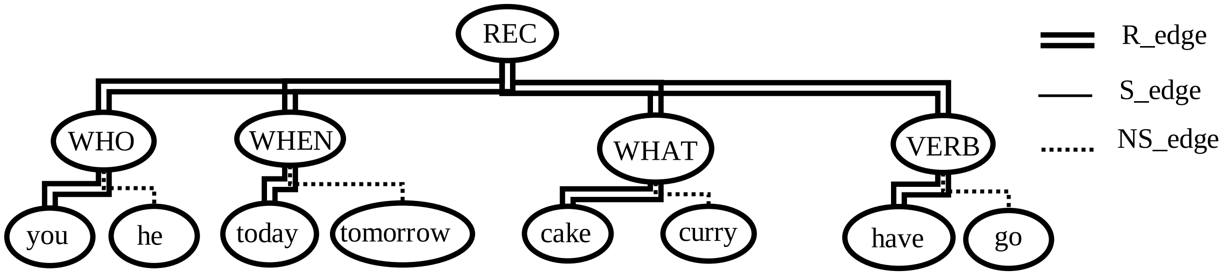


Fig. 2. A part of the recommendation tree of "How about having a cake today?"

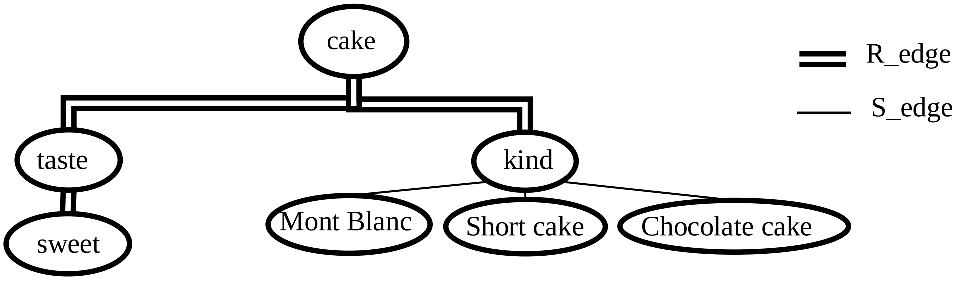


Fig. 3. The subtree of node "cake"

III. An algorithm of deriving intentions

An algorithm to be proposed here derives the intention of node "REC" from intentions of RC nodes by tracing a recommendation tree. Suppose that there is the acceptance intention of RC node "curry" for a recommendation "How about having a cake today?" in Fig. 2. Then, the intention of node "REC" is derived to the negative intention. Before proposing the algorithm, the following definitions are prepared.

- Definition

Suppose that NODE[*x*] is a node for string *x*. Let EDGE[NODE[*x*],NODE[*y*]] be the kind of edges (*NC-edge*, *S_edge*, and *NS_edge*) between NODE[*x*] and NODE[*y*]. Let INTENTION[NODE[*x*]] be the intention of NODE[*x*] which has one of three kinds intentions: *acceptance*, *rejection*, and *no_information* in this algorithm. *No_information* means that a node doesn't have any intentions. All intentions of nodes are initialized to *no_information*. PARENT(NODE[*x*]) represents the parent node of NODE[*x*]. SIBLING(NODE[*x*]) returns the set of sibling nodes of NODE[*x*].

- Rejection intentions of PARENT(NODE[*x*])

INTENTION[PARENT[NODE[*x*]]] is *rejection* if REJECTION(NODE[*x*]) is true. It is computed by (1)-(4), where "∨" and "∧" means logical disjunction and logical conjunction, respectively.

$$\begin{aligned}\text{REJECTION}(\text{NODE}[x]) &= \text{REJECTION1}(\text{NODE}[x])\\ &\lor \text{REJECTION2}(\text{NODE}[x])\\ &\lor \text{REJECTION3}(\text{NODE}[x])\end{aligned} \tag{1}$$

$$\begin{aligned}&\text{REJECTION1}(\text{NODE}[x]) =\\ &\exists x(\text{INTENTION}[\text{NODE}[x]] = \textit{rejection}\\ &\land\ \text{EDGE}[\text{NODE}[x], \text{PARENT}(\text{NODE}[x])] =\\ &\textit{R\_edge})\end{aligned} \tag{2}$$

$$\begin{aligned}&\text{REJECTION2}(\text{NODE}[x]) = \exists x(\forall\\ &\text{SIBLING}(\text{NODE}[x])(\text{INTENTION}[\text{SIBLING}(\text{NODE}\\ &[x])] = \textit{rejection}\\ &\land\ \text{EDGE}[\text{PARENT}(\text{NODE}[x]),\\ &\text{SIBLING}(\text{NODE}[x])] = \textit{S\_edge})\\ &\land\ (\text{INTENTION}[\text{NODE}[x]] = \textit{rejection}\ \land\\ &\text{EDGE}[\text{PARENT}(\text{NODE}[x]), \text{NODE}[x]] = \textit{S\_edge}))\end{aligned} \tag{3}$$

$$\begin{aligned}&\text{REJECTION3}(\text{NODE}[x]) =\\ &\exists x(\text{INTENTION}[\text{NODE}[x]] = \textit{acceptance}\\ &\land\ \text{EDGE}[\text{NODE}[x], \text{PARENT}(\text{NODE}[x])] =\\ &\textit{NS\_edge})\end{aligned} \tag{4}$$

Suppose that users refuse NODE["cake"], and then INTENTION[NODE["cake"]] is *rejection*. In Fig. 2, PARENT(NODE["cake"]) is NODE["WHAT"], and EDGE[NODE["cake"],NODE["WHAT"]] is equal to *R_edge*. For REJECTION 1, INTENTION[NODE["WHAT"]] is *rejection.*

Next, suppose that users refuse NODE[*x*] for all *x* such that *x* is "chocolate cake", "short cake", and "Mont Blanc". Then, INTENTION[NODE[*x*]] is *rejection*. In Figs. 2 and 3, PARENT(NODE[*x*]) is NODE["kind"], and EDGE[NODE["kind"],NODE[*x*]] is *S_edge*. For REJECTION 2, INTENTION[NODE["kind"]] is *rejection*.

Finally, suppose that users accept NODE["curry"], and then INTENTION[NODE["curry"]] is *acceptance*. In Fig. 2, PARENT[NODE["curry"]] is NODE["WHAT"], and

EDGE[NODE["curry"],NODE["WHAT"]] is equal to *NS_edge*. For REJECTION 3, INTENTION[NODE["WHAT"]] is *rejection*.

- Acceptance intentions of PARENT(NODE[*x*])

INTENTION[PARENT(NODE[*x*])] is *acceptance* if ACCEPTANCE(NODE[*x*]) is true. It is computed by (5)-(7).

$$\begin{aligned}&\text{ACCEPTANCE}(\text{NODE}[x]) = \\ &\text{ACCEPTANCE1}(\text{NODE}[x]) \\ &\vee\ \text{ACCEPTANCE2}(\text{NODE}[x])\end{aligned} \tag{5}$$

$$\begin{aligned}&\text{ACCEPTANCE1}(\text{NODE}[x]) = \\ &\exists x(\text{INTENTION}[\text{NODE}[x]] = \textit{acceptance} \\ &\wedge\ \text{EDGE}[\text{NODE}[x],\ \text{PARENT}(\text{NODE}[x])] = \\ &(\textit{R\_edge} \vee \textit{S\_edge}))\end{aligned} \tag{6}$$

$$\begin{aligned}&\text{ACCEPTANCE2}(\text{NODE}[x]) = \\ &\exists x(\text{INTENTION}[\text{NODE}[x] = \textit{rejection} \\ &\wedge\ \text{EDGE}[\text{PARENT}(\text{NODE}[x]),\ \text{NODE}[x]] = \textit{S\_edge} \\ &\wedge\ \exists\,\text{SIBLING}(\text{NODE}[x]) \\ &(\text{INTENTION}[\text{SIBLING}(\text{NODE}[x])] \neq \textit{rejection} \\ &\wedge\ \text{EDGE}[\text{PARENT}(\text{NODE}[x]), \\ &\text{SIBLING}(\text{NODE}[x])] = \textit{S\_edge}))\end{aligned} \tag{7}$$

Suppose that users accept NODE["cake"], and then INTENTION[NODE["cake"]] is *acceptance*. In Fig. 2, PARENT(NODE["cake"]) is NODE["WHAT"], and EDGE[NODE["cake"],NODE["WHAT"]] is equal to *R_edge*. For ACCEPTANCE 1, INTENTION[NODE["WHAT"]] is *acceptance*.

Next, suppose that users accept NODE["chocolate cake"], and then INTENTION[NODE["chocolate cake"]] is *acceptance*. In Fig. 2, PARENT(NODE["chocolate cake"]) is NODE["WHAT"], and EDGE[NODE["chocolate cake"],NODE["WHAT"]] is equal to *S_edge*. For ACCEPTANCE 1, INTENTION[NODE["WHAT"]] is *acceptance*.

Finally, suppose that users reject NODE["chocolate cake"] and don't reject NODE["short cake"]. Then, INTENTION[NODE["chocolate cake"]] is *rejection*, and INTENTION[NODE["short cake"]] is *no_information* or *acceptance*. In Fig. 3, PARENT(NODE[*x*]) for *x*="chocolate cake" and *x*="short cake" is NODE["kind"], and EDGE[NODE["kind"],NODE[*x*]] is *S_edge*. For ACCEPTANCE 2, INTENTION[NODE["kind"]] is *acceptance*.

By using above definitions, the proposed algorithm is defined as below.

- An algorithm of deriving intentions

Input: ANSWER_NODE[] and ANSWER_INTENTION[]

ANSWER_NODE[] is a list of strings for nodes accepted or rejected by answers. ANSWER_INTENTION[] is a list of intentions for elements in ANSWER_NODE[]. Indexes of ANSWER_INTENTION[] are elements in ANSWER_NODE[]. For the answer "I like curries", ANSWER_NODE[] is {"curry"} and ANSWER_INTENTION["curry"] is {"*acceptance*"}, respectively.

Output: INTENTION[NODE["REC"]]

Method:

```
 for i=1 to n do/*n is the number of elements in ANSWER_NODE[]*/
    INTENTION[NODE[ANSWER_NODE[i]]=ANSWER_INTENTION[ANSWER_NODE[i]]
    target_node = NODE[ANSWER_NODE[i]]
  while target_node $\neq$ NODE["REC"] do
      if REJECTION(target_node) is true then
      INTENTION[ PARENT(target_node)] = rejection
    else if ACCEPTANCE(target_node) is true then
      INTENTION[ PARENT(target_node)] = acceptance
    endif
    target_node = PARENT(target_node)]
  endwhile
  if INTENTION[ NODE["REC"]] is rejection then
    INTENTION[ NODE["REC"]] = negative
    break
  else if INTENTION[ NODE["REC"]] is acceptance then
    INTENTION[ NODE["REC"]] = affirmative
  endif
endfor
```

End of Algorithm

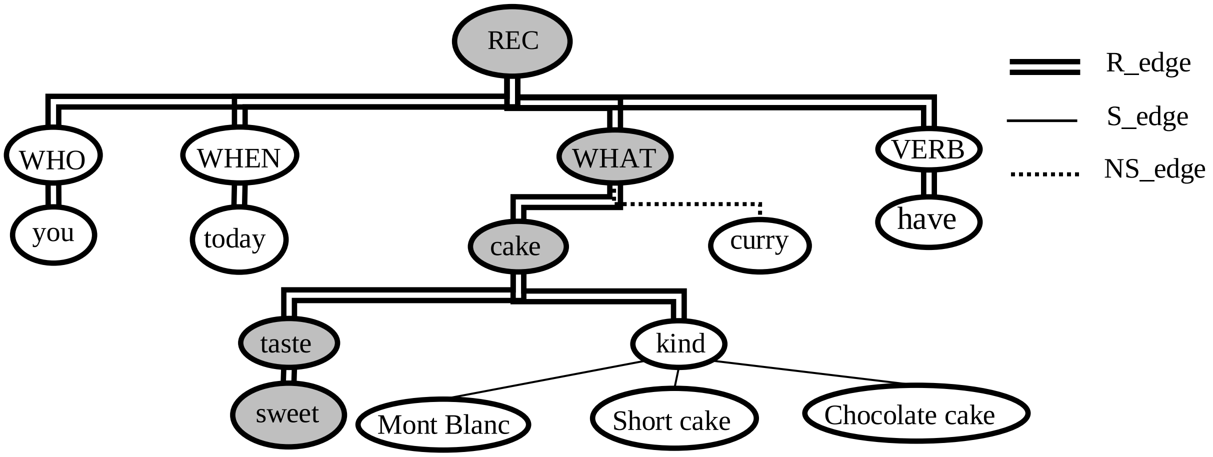


Fig. 4. The derivation process of the answer "I like something sweet."

In case of a recommendation "How about having a cake today?", examples of derivations from answers "I like something sweet." (accepting the node with *R_edge*), "I hate something sweet." (rejecting the node with *R_edge*), "I like curries." (accepting the node with *NS_edge*), and "I dislike short cakes" (rejecting the node with *S_edge*) are as follows:

Example 3.1

For an answer "I like something sweet.", ANSWER_NODE[] is {"sweet"} and ANSWER_INTENTION["sweet"] is {"*acceptance*"}.

By tracing NODE[*x*] for all *x* such that *x* is "sweet", "taste", "cake", "WHAT", and "REC", INTENTION[NODE["REC"]] is become *affirmative*. Fig. 4 shows intentions of nodes. Gray-shaded circles represent nodes with the intention of *acceptance*.

Example 3.2

For an answer "I hate something sweet.", ANSWER_NODE[] is {"sweet"} and ANSWER_INTENTION["sweet"] is {"*rejection*"}.

By tracing NODE[*x*] for all *x* such that *x* is "sweet", "taste", "cake", "WHAT", and "REC", INTENTION[NODE["REC"]] is become *negative*. Fig. 5 shows intentions of nodes. Black-shaded circles represent nodes with the intention of *rejection*.

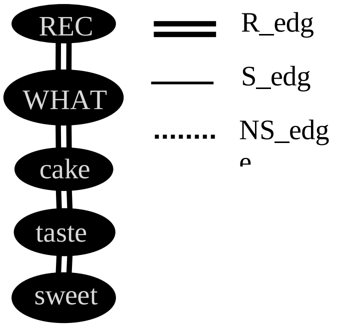


Fig. 5. The derivation process of the answer "I hate something sweet."

Example 3.3

For an answer "I like curries.", ANSWER_NODE[] is {"curry"} and ANSWER_INTENTION["curry"] is {"*acceptance*"}.

By tracing NODE[*x*] for all *x* such that *x* is "curry", "WHAT", and "REC", INTENTION[NODE["REC"]] is become *negative*. Fig. 6 shows intentions of nodes. Gray-shaded circles and black-shaded circles represent nodes with intentions of *acceptance* and *rejection*, respectively.

Example 3.4

For an answer "I dislike short cakes", ANSWER_NODE[] is {"short cake"} and ANSWER_INTENTION["short cake"] is {"*rejection*"}.

By tracing NODE[*x*] for all *x* such that *x* is "short cake", "kind", "cake", "WHAT", and "REC", INTENTION[NODE["REC"]] is become *affirmative*. Fig. 7 shows intentions of nodes. Gray-shaded circles and black-shaded circles represent nodes with intentions of *acceptance* and *rejection*, respectively.

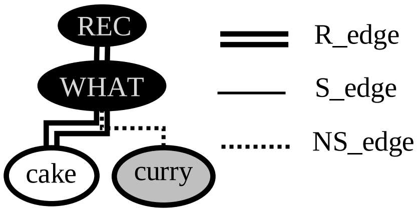


Fig. 6. The derivation process of the answer "I like curries."

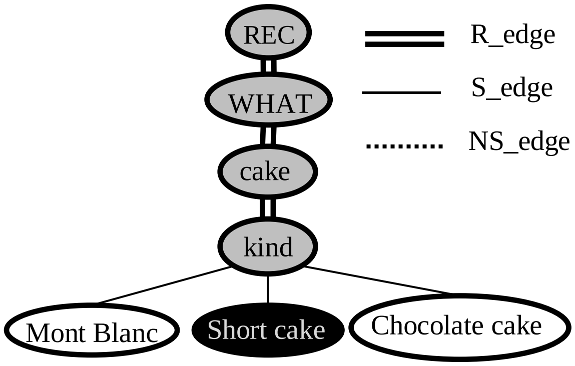


Fig. 7. The derivation process of the answer "I dislike short cakes."

IV. Experiments

A. Knowledge for experiments

In daily lives, it is common to recommend foods including cakes and Japanese noodles, and movies. In this experiment, the following three recommendations are assumed, where "Resident Evil" is a title of a movie:

- Recommendation 1: "How about having a cake today?"
- Recommendation 2: "How about having a Japanese noodle today?"
- Recommendation 3: "How about going to the movie, Resident Evil?"

In order to determine intentions from answers to them, a recommendation tree is needed. This experiment constructs the tree from closed corpora which have 500 answers for each recommendation. Answers are collected by four undergraduate students. From corpora, RC nodes and the recommendation tree are defined by discussions with these students. Examples of RC nodes with concept nodes are presented in Table 2. For RC nodes "cake", "Japanese noodle", and "Resident Evil", more detailed descendant nodes are constructed. Total numbers of descendant nodes are 236 nodes. Examples of descendant nodes for each RC node are presented in Table 3. In Table 3, *R_edge* and *S_edge* between a node and a parent node show R and S, respectively.

TABLE II. Examples of RC nodes with concept nodes

Concept nodes	RC nodes	Numbers
WHO	You, He, She	7
WHEN	Now, Today, Tomorrow	7
WHAT	Cake, Japanese noodle, Resident Evil	51
VERB	Have, See	10

TABLE III. Examples of descendant nodes of nodes"cake", "Japanese noodle", and "Resident Evil"

Parent	Child	Grandchild
Cake	Genre[R]	Sweets[R], Dessert[R], Confectionery[R]
	Taste[R]	Sweet[R]
	Kind[R]	Short cake[S], Mont Blanc[S], Mille-feuille[S]
	Ingredient[R]	Flour[R], Sugar[R], Egg[R] Butter[S], Apple[S], Strawberry[S], Banana[S]
Japanese noodle	Genre[R]	Noodles[R], Food[R]
	Taste[R]	Spicy[S], Light[S], Salty[S]
	Kind of soup[R]	Miso[S], Soy[S], Salt[S]
	Kind of noodle[R]	Crimp[S], Straight[S], Thin[S], Thick[S] Flour[R]
	Ingredient[R]	Garlic[S], Bean sprouts[S], Onion[S], Sesame seeds[S]
	Size[R]	Large[S], Medium[S], Small[S]
Resident Evil	Screen type[R]	Caption[S], Dub[S], 3D[S]
	Genre of films[R]	Horror[R], Action[R]

(R and S means required and selectable)

B. Knowledge for experiments

In order to evaluate the accuracy of the proposed method, two experiments for closed and open tests are carried out. The closed test uses corpora for constructing the recommendation tree with 500 answers for each recommendation. Open tests uses corpora with 100 answers for each recommendation such as the appendix of this paper. These corpora are collected by ten undergraduate students who don't accumulate closed corpora, and they make ten answers to each recommendation without restriction of responses.

The traditional method proposed by Yoshie et al. [21] is used as a comparative method. Table 4 shows results on the closed test of the proposed method. Tables 5 and 6 show results on open tests for the proposed method and the

TABLE IV. RESULTS ON THE CLOSED TEST OF THE PROPOSED METHOD

	Correct sentences	Total sentences	Correct rates (%)
Recommendation 1	472	509	92.7
Recommendation 2	512	556	92.1
Recommendation 3	476	506	94.1

TABLE V. RESULTS ON THE OPEN TEST OF THE PROPOSED METHOD

	Correct sentences	Total sentences	Correct rates (%)
Recommendation 1	81	100	81.0
Recommendation 2	83	100	83.0
Recommendation 3	84	100	84.0

TABLE VI. RESULTS ON THE OPEN TEST OF THE COMPARATIVE METHOD

	Correct sentences	Total sentences	Correct rates (%)
Recommendation 1	38	100	38.0
Recommendation 2	40	100	40.0
Recommendation 3	34	100	34.0

Comparative methods, respectively. In tables 4, 5, and 6, correct rates mean percentages of correct sentences in total sentences.

From Table 4, it is verified that correct rates of the proposed method becomes high for the closed tests. From Tables 5 and 6, all accuracies of the proposed method are about 40 points higher than the comparative method in open tests of recommendations 1, 2, and 3.

In the open test, problems of the proposed method are misclassifications of complex sentences and the lack of knowledge.

- Misclassifications of complex sentences

The example of the complex sentence is “I’ve had enough, therefore I choose a small dish.” which has two sentences, “I’ve had enough” and “I choose a small dish”. The intention of the first sentence is negative because the sentence rejects R_RC,”have”. The intention of the second sentence is affirmative because the sentence accepts S_RC, “small”. The intention of the sentence is affirmative. However, the proposed method produces a negative intention because the negative intention is prior to the affirmative one.

The way to solve the problem is to consider conjunctions and give priority to the intention estimated from a backward-sentence.

- The lack of knowledge

A part of misclassifications of the proposed method are occurred in sentences which include non-defined nodes. For example, the proposed method misclassifies the sentence, “I have to go to a piano lesson.”, because there is no node “piano lesson” in the child nodes of node “WHAT”. This problem can be solved by introducing knowledge of daily lives.

V. CONCLUSIONS

This paper has proposed a method of determining affirmative and negative intentions from indirect speech acts. In the proposed method, a recommendation tree has been defined and an algorithm of deriving intentions of indirect speech acts by the tree is proposed.

The tree consists of nodes and edges corresponding to the three kinds of RC: R_RC, S_RC, and NS_RC. The root node indicates the recommendation and has four concept nodes as child nodes. Concept nodes have RC nodes as child nodes. The deriving algorithm determines affirmative and negative intentions of indirect speech acts by tracing the trees.

From experimental results for three kinds of open tests, all accuracies of the proposed method are about 40 points higher than the traditional method.

VI. FUTURE WORK

Future works are to improve misclassifications of complex sentences of an acceptance sentence and a rejection sentence, and to construct invitational knowledge of daily lives.

REFERENCES

[1] A. Levi, O. Monkryn, C. Diot, and N. Traft, "Finding a needle in a haystack of reviews: cold start context-based hotel recommender system demo." Proceedings of the sixth ACM conference on Recommender systems. ACM, 2012.

[2] P. Lops, G. Marco, and S. Giovanni, "Content-based recommender systems: State of the art and trends." Recommender Systems Handbook. Springer US, 73-105, 2011.

[3] B. J. Han, S. Rho, S. Jun, and E. Hwang, "Music emotion classification and context-based music recommendation." Multimedia Tools and Applications 47.3, 433-460, 2010.

[4] P. Johansson, “Natural language interaction in personalized EPGs”, InProc. of Workshop notes from the 3rd International Workshop on Personalization of Future TV, Johnstown, Pennsylvania, USA , 27-31, 2003.

[5] P. Johansson, “Madfilm-a multimodal approach to handle search and organization in a movie recommendation system”, In Proceedings of the 1st Nordic Symposium on Multimodal Communication , 53-65, 2003.

[6] T. Misu, and T. Kawahara, “Speech-based interactive information guidance system using question-answering technique”, In Acoustics, Speech and Signal Processing, 2007. ICASSP 2007. , 4, 145-148.

[7] H. Shimazu, “ExpertClerk: navigating shoppers' buying process with the combination of asking and proposing”, In Proceedings of the 17th international joint conference on Artificial intelligence, 2, 1443-1448, 2001.

[8] H. Shimazu, “ExpertClerk: A Conversational Case-Based Reasoning Tool forDeveloping Salesclerk Agents in E-Commerce Webshops”, Artificial Intelligence Review, 18(3-4), 223-244, 2002.

[9] C. A. Thompson, M. H. Goeker, and P. Langley, “A personalized system for conversational recommendations”, J. Artif. Intell. Res. (JAIR), 21, 393-428, 2004.

[10] R. Fernandez, and R. W. Picard, “Dialog act classification from prosodic features using support vector machines”, In Speech Prosody 2002, International Conference, 291-294.

[11] L. Levin, A. Thymé-Gobbel, A. Lavie, K. Ries and K. Zechner, “A discourse coding scheme for conversational Spanish”, In International Conference on Speech and Language Processing, 1998.

[12] D. Surendran, and G. A. Levow, “Dialog act tagging with support vector machines and hidden Markov models”, In Proceedings of Interspeech , 2006, 1950-1953.

[13] A. H. Anderson, M. Bader, E. G. Bard, E. Boyle, G. Doherty, S. Garrod, S. Isard. J. Kowtko, J. McAllister, J. Miller, C. Sotillo, H. S. Thompson, and R. Weinert, "The HCRC map task corpus", Language and Speech,34(4), 351-366, 1991.

[14] J. Carletta, S. Isard, G. Doherty-Sneddon, A. Isard, J. C. Kowtko, and A. H. Anderson, "The reliability of a dialogue structure coding scheme", Computational Linguistics, 23(1), 13-31, 1997.

[15] A. Stolcke, K. Ries, N. Coccaro, E. Shriberg, R. Bates, D. Jurafsky, P. Taylor, R. Martin, C. V. Ess-Dykema, and M. Meteer, "Dialogue act modeling for automatic tagging and recognition of conversational speech", Computational linguistics, 26(3), 339-373, 2000.

[16] M. Core, and J. Allen, "Coding dialogs with the DAMSL annotation scheme", In AAAI fall symposium on communicative action in humans and machines, 28-35, 1997.

[17] S. Ravi, and J. Kim, "Profiling student interactions in threaded discussions with speech act classifiers", Frontiers in Artificial Intelligence and Applications, 357-364, 2007.

[18] K. Mera, T. Ichimura, and T. Yamashita, "Analysis of User Communicative Intention from Affirmative/Negative Elements by Fuzzy Reasoning and Its Application to WWW-based Health Service System for Elderly", In Proc. of the 6th Intl. Conf. on Soft Computing (IIZUKA2000) , 971-976.

[19] J. Kitamura, Y. Watanabe, Y. Sekiguchi, and Y. Suzuki, "An Extraction and Processing Method of User's Denial Utterance for a Speech Dialog Device", Transactions of Information Processing Society of Japan, 46(7), 1789-1796, 2005. (in Japanese)

[20] K. Mera, "Analyzing affirmative/negative intention from plural sentences", Proc. KES'01, 1222-1227.

[21] M. Yoshie, K. Mera, T. Ichimura, T. Yamashita, T. Aizawa, and K. Yoshida, "Analysis of affirmative/negative intentions of the answers to yes-no questions and its application to a web-based interface", Journal of Japan Society for Fuzzy Theory and Systems, 14(4), 393-403, 2002.

Appendix

Examples of answers in Open corpora are shown as follows:

[Answers to recommendation 1, "How about having a cake today?"]

Answers	Intention
Let's go now.	Affirmative
I can't get enough of sweet food.	Affirmative
I have a Mont Blanc.	Affirmative
I dislike short cakes.	Affirmative
Oh goody!	Affirmative
My tummy is full.	Negative
Maybe another time.	Negative
I don't like cakes.	Negative
I can't have it because I have a piano lesson now.	Negative
I don't have it because I have egg allergies.	Negative

[Answers to recommendation 2, "How about having a Japanese noodle today?"]

Answers	Intention
Yes, I want to have it.	Affirmative
Let's go to a low price Japanese noodle shop.	Affirmative
I want to have 3 bowls.	Affirmative
I want to have a miso-flavored noodle.	Affirmative
O.K.	Affirmative
I got tired of it.	Negative
Shall we have other foods?	Negative
I can't have it because I feel bad.	Negative
I don't go to have it.	Negative
I can't have noodles.	Negative

[Answers to recommendation 2, "How about going to the movie, Resident Evil?"]

Answers	Intention
I want to watch it with 3D scenography.	Affirmative
I like action movies.	Affirmative
It seems pleasant.	Affirmative
I want to go if the ticket price is discounted.	Affirmative
I like movies very much.	Affirmative
I want to watch animation movies.	Negative
I don't want to watch it.	Negative
I watch it by a rental video.	Negative
I dislike something terrible.	Negative
I'm not interested it.	Negative

Performance Analysis of Keccak *f*-[1600]

performance based on storage space requirements

Ananya Chowdhury
Department of Information Technology
Jadavpur University
Kolkata, West Bengal, India

Utpal Kumar Ray
Department of Information Technology
Jadavpur University
Kolkata, West Bengal, India

***Abstract*—Keccak is the latest Hash Function selected as the winner of NIST Hash Function Competition. SHA-3 is not meant to replace SHA-2 as no significant attacks on SHA-2 have been demonstrated. But it is designed in response to the need to find an alternative and dissimilar construct for Cryptographic Hash that is more fortified to attacks. In this paper we have tried to depict an analysis of the software implementation of Keccak-*f*[1600] based on the disk space utilization and time required to compute digest of desired sizes.**

Keywords—Sponge Construction; State; Rounds; Bitrate(r); Capacity(c); Diversifier(d); Plane; Slice; Sheet; Row; Column; Lane; Bit

I. Background

Distributed Computing and Network Communication has revolutionized the face of modern computing. But it brings with it serious security concerns like verifying the integrity and authenticity of the transmitted data. The sender and the receiver communicating over an insecure channel essentially require a method by which the information transmitted by the sender can be easily authenticated by the receiver as “unmodified” or authentic. To achieve this, technique called “Hashing” is employed which relies on a family of Hash Functions. Keccak is one such Hash Function which is selected as the winner of NIST Hash Function Competition.

II. Introduction

The state of Keccak- *f* [1600] is organized as a three-dimensional array [2], which suggests several ways to partition the bits. The naming conventions as suggested by the authors are described in detail in the subsequent sections. While this is an optimal choice on software platforms actually offering 64-bit operations, the bit interleaving technique allows efficient implementations on systems with smaller word sizes and can also be used to target compact hardware circuits. In its simplest form, namely factor-2 interleaving, it splits the odd and even bits of each lane. The state of Keccak- *f* [1600] is then represented as 50 words of 32 bits.

III. Sponge Function

A. *What is a Sponge Function ?*

In the context of cryptography, the Sponge Construction[2] is a mode of operation, based on a fixed-length permutation (or transformation) and on a padding rule, which builds a function mapping variable-length input to variable-length output. It takes as input an element of $(Z_2)^*$, i.e., a binary string of any length, and returns a binary string with any requested length, i.e., an element of $(Z_2)^n$ with n a user-supplied value. It operates on a finite state by iteratively applying the inner permutation to it, interleaved with the entry of input or the retrieval of output.

B. *Working Principle of Sponge Construction*

The sponge construction[2] is a simple iterated construction for building a function F, with variable-length input and arbitrary length output based on a fixed-length permutation (or transformation) f, operating on a fixed number, b of bits. Here b is called the width. The sponge construction operates on a state of **b =(r + c)** bits. The value r is called bit-rate and c is called capacity.

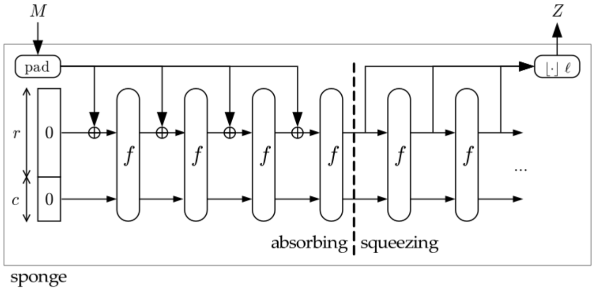


Fig. 1. Sponge Construction

First, the input string is padded with a reversible padding rule and cut into blocks of r bits. Then the b bits of the state are initialized to zero and the sponge construction proceeds in two phases:

- In the absorbing phase, the r-bit input blocks are XOR ed into the first r bits of the state, interleaved with applications of the function f. When all input blocks are processed, the sponge construction switches to the squeezing phase.
- In the squeezing phase, the first r bits of the state are returned as output blocks, interleaved with applications of the function f. The number of output blocks is chosen at will by the user.

The sponge construction uses r +c bits of state, of which r are updated with message bits between each application of Keccak- *f* during the absorbing phase and output during the squeezing phase. The remaining c bits are not directly affected by message bits, nor are they taken as output.

IV. Naming Conventions

The Keccak naming conventions are as follows:

- **State and Rounds [3]:** Keccak consists of a set of 7 permutations and is denoted as Keccak - *f* [b], where b {25, 50, 100, 200, 400, 800, 1600} is the width of the permutation. The state of Keccak- *f* [1600] is organized as a three-dimensional array, which suggests several ways to partition the bits. The state of Keccak- *f* [1600] can be expressed as 25 lanes of 64 bits each. These Keccak-*f* permutations are iterated constructions consisting of a sequence of almost identical rounds. The number of rounds n_r depends on the permutation width, and is given by

∈

$\mathbf{n_r = 12 + 2\ell}$, where $\mathbf{2^\ell = b/25}$ **.**

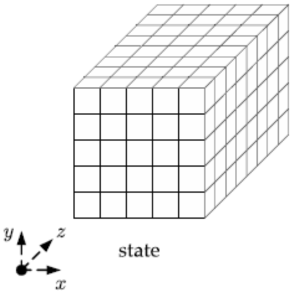


Fig. 2. State

∈

- Bit-rate(r), Capacity(c) and Diversifier(d) [3]:

The sum b = r + c determines the width of the Keccak-*f* permutation used in the Sponge Construction where b {25, 50, 100, 200, 400, 800, 1600}. The diversifier value satisfies 0<=d< 256.

The default bitrate r = 1024 is a power 10 of 2 to ease data alignment and the resulting capacity is c = 1600−1024 = 576. The default value for the diversifier d is 0.

The purpose of the diversifier is to provide diversification, i.e., two instances of Keccak with two different values of d behave as two independent ha∈ı functions (even with same values of r and c).

- **Plane [3]:** A plane is a set of 5w bits with constant y coordinate.

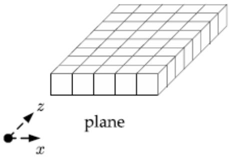


Fig. 3. Plane

- **Slice [3]:** A slice is a set of 25 bits with constant z coordinate.

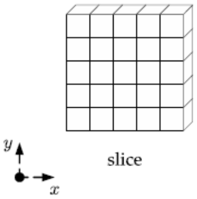


Fig. 4. Slice

- **Sheet [3]:** A sheet is a set of 5w bits with constant x coordinate.

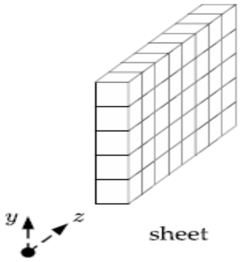


Fig. 5. Sheet

- **Row [3]:** A row [15] is a set of 5 bits with constant y and z coordinates.

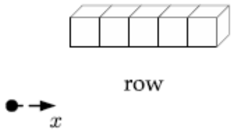


Fig. 6. Row

- **Column [3]:** A column [15] is a set of 5 bits with constant x and z coordinates.

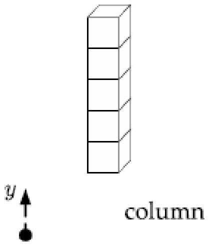


Fig. 7. Column

- **Lane [3]:** A lane [15] is a set of w bits with constant x and y coordinates.

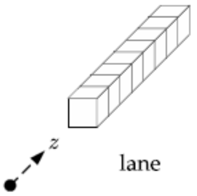


Fig. 8. Lane

- **Bit [3]:** A particular w bit [15] is referred to as bit.

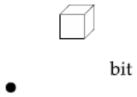


Fig. 9. Bit

V. Specification Summary of Keccak

The specification of Keccak-*f* [1600] is as follows.

```
Keccak-f[b] (A) {
 forall i in 0…$n_r$-1
A = Round[b] (A, RC[i])
return A
}

Round[b] (A, RC) {

 θ step
C[x] = A[x, 0] xor A[x, 1] xor A[x, 2] xor A[x, 3] xor A[x, 4],
forall x in 0…4
D[x] = C[x-1] xor rot(C[x+1], 1),
forall x in 0…4
A[x, y] = A[x, y] xor D[x],
forall (x, y) in (0…4, 0…4)

ρ and π steps
B[y, 2*x+3*y] = rot (A[x, y], r[x, y]),
forall (x, y) in (0…4, 0…4)

χ step
A[x, y] = B[x, y] xor ((not B[x+1, y]) and B[x+2, y]),
forall (x, y) in (0…4, 0…4)

ι step
A [0, 0] = A [0, 0] xor RC

return A
}
```

All the operations on the indices are done modulo 5. A denotes the complete permutation state array, and A[x, y] denotes a particular lane in that state. B[x, y], C[x], D[x] are intermediate variables.

The constants r[x, y] are the rotation offsets, while RC[i] are the round constants. rot (W, r) is the usual bitwise cyclic shift operation, moving bit at position i into position i+r (modulo the lane size).

A Keccak- *f* round consists of a sequence of invertible steps each operating on the state, organized as an array of 5 X 5 lanes, each of length w {1,2 4, 8, 16, 32, 64} (b = 25w). Therefore b {25, 50, 100, 200, 400, 800, 1600}. When implemented on a 64-bit processor, a la∈e of Keccak- *f* [1600] can be represented as a 6∈-bitCPU word. Here not denotes the bitwise exclusive OR, NOT the bitwise complement and AND the bitwise AND operation.

We obtain the Keccak[r, c] sponge function, with parameters capacity, c and bit-rate, r if we apply the sponge construction to Keccak - *f* [r + c] and perform specific padding on the message input.

In the pseudo-code below, S denotes the state as an array of lanes. The padded message P is organised as an array of blocks P_i, themselves organized as arrays of lanes. The || operator denotes the usual byte string concatenation.

```
Keccak[r,c](M) {

 Initialization and Padding

                    S[x, y] =0,
forall (x, y) in (0…4, 0…4)

                    P = M || 0x01 || 0x00 || … || 0x00

                    P = P xor (0x00 || … || 0x00 || 0x80)

Absorbing Phase

                    forall block Pi in P

                    S[x, y] = S[x, y] xor Pi[x+5*y],
forall (x, y) such that x+5*y < r/w

                    S = Keccak-f[r+c](S)

 Squeezing Phase

                    Z = empty string

                    while output is requested

                    Z = Z || S[x, y],
forall (x, y) such that x+5*y < r/w

                    S = Keccak-f[r+c](S)

Return Z

}
```

VI. Experimental Setup

This section describes the inputs, outputs, experimental results and graphical analysis after implementation of Keccak-*f* [1600] under the laboratory experimental setup of University. All the experiments are performed on the following hardware and software platform and the experimental results are recorded with the best possible precision and accuracy under the laboratory experimental setup.

1) *Hardware Configuration:*

- Intel® Core(TM)2 Quad CPU

  Q8400 @2.66GHz, 2.65GHz
  3.25GB RAM
  Physical Address Extension

2) *Operating System:*

- Fedora release 11(Leonidas)

3) Software Configuration:

- Language used is C
- Compiler: gcc (GCC) 4.4.0 20090506(Red Hat 4.4.0-4)

Our chief objective is to analyze the performance of Keccak-*f* [1600] with respect to the time required to compute digests of various sizes and the disk space required by the output files of different.

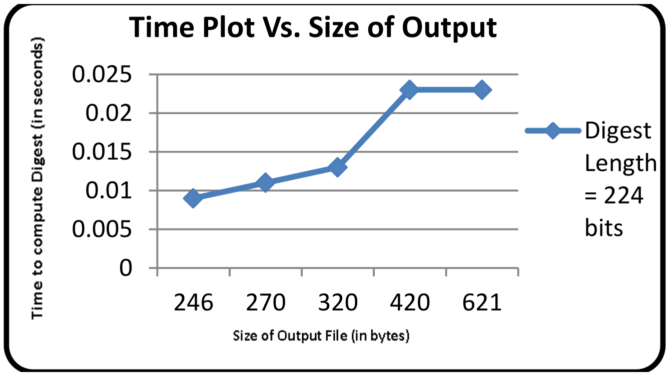


Fig. 10. Time Taken to Compute Digest Vs. Size of Output File Plot for Digest of Size = 224bits

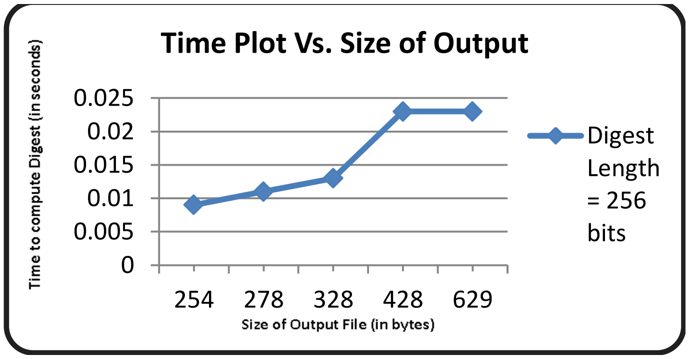


Fig. 11. Time Taken to Compute Digest Vs. Size of Output File Plot for Digest of Size = 256bits

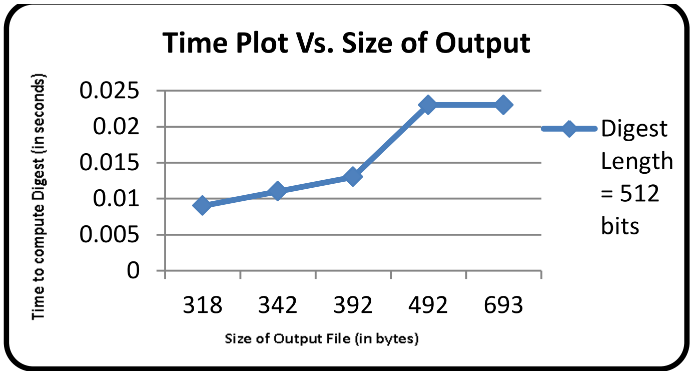


Fig. 12. Time Taken to Compute Digest Vs. Size of Output File Plot for Digest of Size = 512 bits

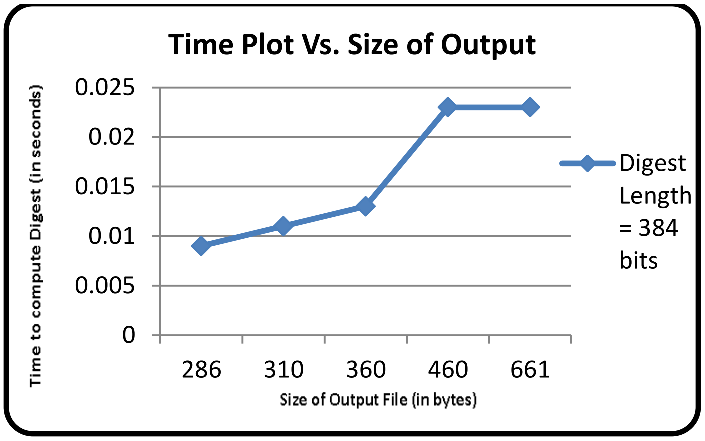


Fig. 13. Time Taken to Compute Digest Vs. Size of Output File Plot for Digest of Size = 384bits

TABLE I. TIME AND SIZE OF OUTPUT

Time to compute Digest(in seconds)	**Size of Output File for different Digest Lengths(in bytes)**			
	Digest Length = 224 bits	***Digest Length = 256 bits***	***Digest Length = 384 bits***	***Digest Length = 512 bits***
0.009	246	254	286	318
0.011	270	278	310	342
0.013	320	328	360	392
0.023	420	428	460	492
0.023	621	629	661	693

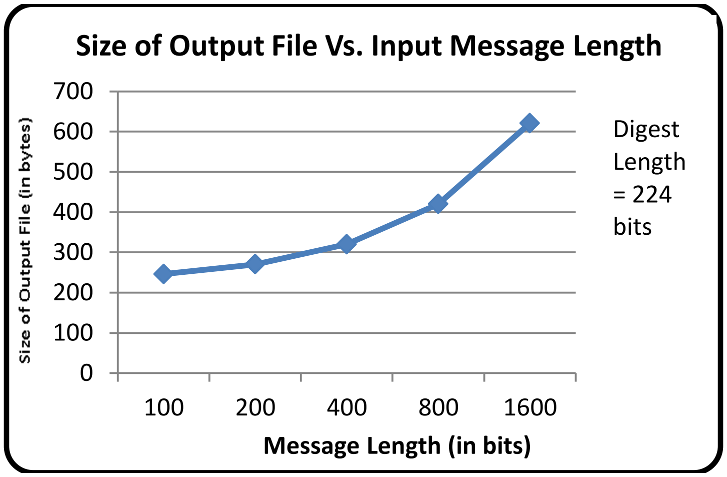


Fig. 14. Size of Output File Vs. Input Message Length Plot for Digest of Size = 224bits

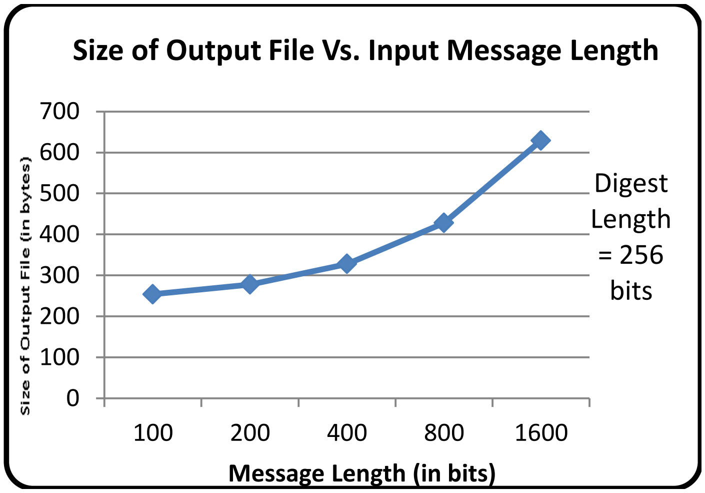


Fig. 15. Size of Output File Vs. Input Message Length Plot for Digest of Size = 256bits

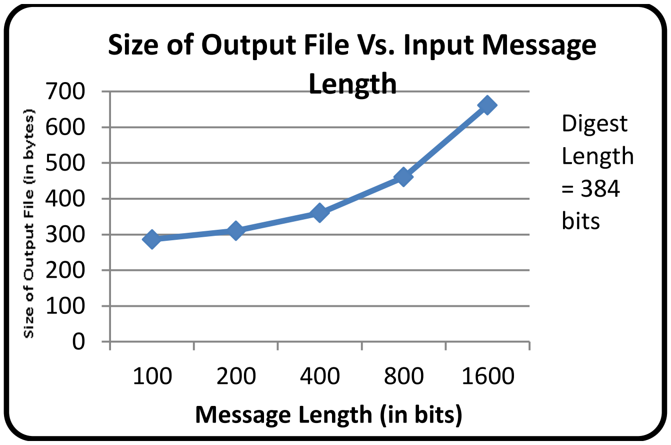


Fig. 16. Size of Output File Vs. Input Message Length Plot for Digest of Size = 384bits

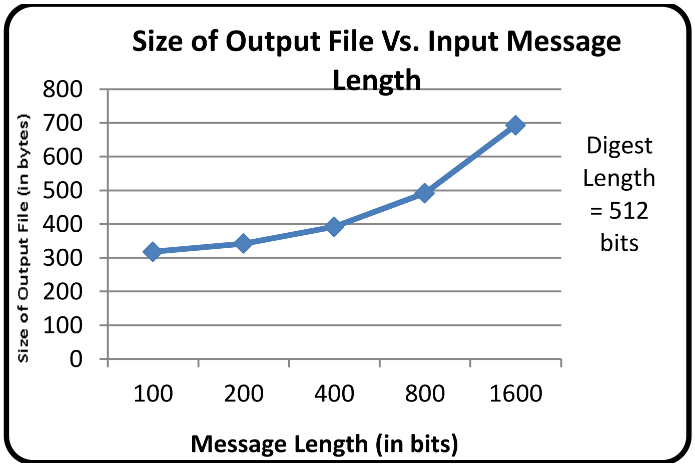


Fig. 17. Size of Output File Vs. Input Message Length Plot for Digest of Size = 512bits

TABLE II. INPUT MESSAGE LENGTH AND SIZE OF OUTPUT

Length of Input Message(in bits)	**Size of Output File for different Digest Lengths(in bytes)**			
	Digest Length = 224 bits	***Digest Length = 256 bits***	***Digest Length = 384 bits***	***Digest Length = 512 bits***
100	246	254	286	318
200	270	278	310	342
400	320	328	360	392
800	420	428	460	492
1600	621	629	661	693

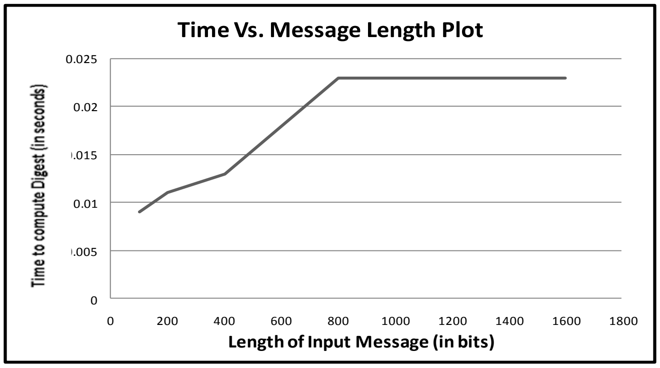


Fig. 18. Time Taken to Compute Digest Vs. Input Message Length Plot for Digest of Size = 512bits

TABLE III. TIME AND MESSAGE LENGTH

Sr. No.	*Length of Input Message(in bits)*	*Time to Compute Digest(in s)*
1	100	0.009
2	200	0.011
3	400	0.013
4	800	0.023
5	1600	0.023

VII. Conclusion And Discussion

It is evident from Fig.10 to Fig. 13, time taken to compute digests of 4 different sizes i.e. 224bits, 256bits, 384bits and 512bits are approximately constant. Initially it rises almost linearly and after a certain point of time it remains constant. This points out a stable behavior of Keccak-f[1600] that almost same time needed to compute digests of different sizes and for large file sizes it is constant.

Thus it will show satisfactory performance for applications that need to compute digests for input of large sizes.

Fig.14 to Fig. 17 depicts how the size of output file grows with an increase in the length of the input message. This observation focuses on the secondary storage requirement of Keccak-*f*[1600]. Keccak-*f*[1600] shows similar graphs for digests of 4 different lengths i.e. 224bits, 256bits, 384bits and 512bits. This shows that Keccak-f[1600] can be conveniently used in devices with limited memory capability like mobile devices.

Fig. 18 shows the time Keccak-f[1600] takes to compute digest of all 4 sizes for different input message lengths. Interestingly for larger input sizes 800 bits and above, the time taken is constant 0.023 s. Thus unlike most other hash functions, the behavior of Keccak-f[1600] is extremely stable and constant for larger input sizes.

As a Sponge Function, Keccak has an arbitrary output length which makes it strikingly different from other well-known Hash Functions which has fixed output length. Keccak does not follow iterated hash function structures like its contemporaries MD5, MD6 etc. The instance of Keccak proposed for SHA-3, Keccak-*f*[1600] make use of a single permutation for all security strengths and this cuts down the implementation cost. Hence Keccak is a robust, flexible, efficient hash algorithm which has a promising future ahead.

VIII. Future Work

Keccak-*f*[1600] can be natively used for hashing, MAC Computation etc. catering to both the needs of fixed-length output and variable length output. In addition it can also be used for symmetric key encryption and random number generation. The arbitrary output length of Keccak makes it suitable for tree hashing. Tree hashing has the power to exploit the advantages of parallel processing for substantially large inputs. This makes Keccak one of most suitable candidate for multi-core processor architecture.

Acknowledgment

With great pleasure we would like to express our heartfelt gratitude to all the staff-members of Department of Information Technology, Jadavpur University, for the help and cooperation.

References

[1] Cryptography and Network Security (Principles and Practices) Fourth Edition by William Stallings.

[2] Guido Bertoni, Joan Daemen, Michael Peeters, Gilles Van Assche “Keccak Implementation Overview”: keccak.noekeon.org_Keccak-implementation-3.2.

[3] Guido Bertoni, Joan Daemen, Michael Peeters, Gilles Van Assche “The Keccak Reference”: keccak.noekeon.org_Keccak-reference-3.0.

Software Ecosystem: Features, Benefits and Challenges

J.V. Joshua, D.O. Alao, S.O. Okolie, O. Awodele
Department of Computer Science, School of Computing and Engineering Sciences, Babcock University, Ilishan-Remo, Ogun State, Nigeria.

Abstract—**Software Ecosystem (SECO) is a new and rapidly evolving phenomenon in the field of software engineering. It is an approach through which many variables can resolve complex relationships among companies in the software industry. SECOs are gaining importance with the advent of the Google Android, Apple iOS, Microsoft and Salesforce.com ecosystems. It is a co-innovation approach by developers, software organisations, and third parties that share common interest in the development of the software technology. There are limited researches that have been done on SECOs hence researchers and practitioners are still eager to elucidate this concept.**

A systematic study was undertaken to present a review of software ecosystems to address the features, benefits and challenges of SECOs.

This paper showed that open source development model and innovative process development were key features of SECOs and the main challenges of SECOs were security, evolution management and infrastructure tools for fostering interaction. Finally SECOs fostered co-innovation, increased attractiveness for new players and decreased costs

Keywords—Software ecosystem; Open source; closed system

I. Introduction

The notion of ecosystems originates from ecology. One definition in Wikipedia defines an ecosystem as a natural unit consisting of all plants, animals and micro-organisms (biotic factors) in an area functioning together with all of the non-living physical (abiotic factors) of the environment.

Although the above is an excellent definition, it is less suitable here and therefore we start from the notion of human ecosystems. A human ecosystem consists of actors, the connections between the actors, the activities by these actors and the transactions along these connections concerning physical or non-physical factors.

Software ecosystems (SECO) refer to the set of businesses and their interrelationships in a common software product or service market [9]. A Software Ecosystem consists of the set of software solutions that enable, support and automate the activities and transactions by the actors in the associated social or business ecosystem and the organizations that provide these solutions [1].

This is an emergent field inspired in concepts from and business and biological ecosystems [14].

Well known examples of communities that may be seen as software ecosystems are Apples iPhone, Microsoft, Google Android, Symbian, Ruby and Eclipse.

Ecosystem concept may refer to a wide range of configurations. Yet, they all involve two fundamental concepts: a network of organisations or actors, and a common interest in the development and use of a central software technology.

The software industry is constantly evolving and is currently undergoing rapid changes. Not only are products and technologies evolving quickly, many innovative companies are experimenting with new business models, leading occasionally to fundamental shifts in entire industry structures and how firms and customers interrelate[17]. Recently, many companies have adopted the strategy of using a platform to attract a mass following of software developers as well as end-users, building entire "software ecosystems" (SECOs) around themselves, even as the business world and the research community are still attempting to get a better understanding of the phenomenon.

This paper explores the main terms under consideration which are the meaning of SECO, identify the main features of Software Ecosystems (SECOs) and finally establish the benefits and challenges of SECOs

II. What is the Problem

In the past few decades, we have witnessed different types of software development methodologies ranging from waterfall, spiral, component, chaos, rapid application development, rational unified process to agile models respectively. Almost all the models mentioned encourage development of software product entirely on the organisation concerned.

The emergent of Software Ecosystem (SECO) development paradigm has brought about co-innovation as a result of different players, however research communities and practitioners are still grasping to understand this concept. Hence this work is aim to expose what is known about software ecosystems (SECOs).

III. Objectives Of The Study

The goal of the study is to carry out a systematic study of software ecosystems in order to present a wider view of what is currently known about software ecosystems

The specific objectives are to:

a) Identify the main features of Software Ecosystems (SECOs).

b) Establish the benefits and challenges of SECOs

IV. Scope Of The Study

It is not easy to study existing Software Ecosystems (SECOs) due to the fact that many SECOs are closed communities and it is hard to get access to information. Therefore, we adopted free open software ecosystems as our subject of studies.

V. Significant Of The Study

The significance of the study is to create awareness about the emergent fields of software ecosystems for research communities and practitioners and to establish research direction for software ecosystems.

VI. Review Of Related Research

Bosch [1] proposed a Software Ecosystem (SECO) taxonomy that identifies nine potential classes of the central software technology as shown in Table1 below, according to classification within two broad dimensions. The first one is the category dimension, which ranges from operating systems to applications, and to end-user programming. The second one is the platform dimension, ranging from desktop to web, and to mobile.

TABLE I. SOFTWARE ECOSYSTEM TAXONOMY

end-user programming	MS Excel, Mathematical,VHDL	Yahoo!Pipes, Microsoft PopFly, Google's mashup editor	none so far
Application	MS Office	SalesForce, eBay, Amazon, Ning	none so far
operating system	MS Windows, Linux, Apple OS X	Google AppEngine, Yahoo developer, Coghead, Bungee Labs	Nokia s60, Palm, Android, iPhone
category platform	Desktop	Web	Mobile

In Software Engineering (SE) community, studies of SECOs were motivated by the software product lines (SPLs) approach aiming at allowing external developers to contribute to hitherto closed platforms [1].

[4], opined that a potential benefit of being a member of a software ecosystem is the opportunity to exploit open innovation an approach derived from open source software (OSS) processes where actors openly collaborate to achieve local and global benefits. External actors and the effort they put into the ecosystem may result in innovations being beneficial not only to themselves (and their customers) but also to the keystone organisation, as this may be a very efficient way of extending and improving the central software technology as well as increasing the number of users.

According to [8] closer relationships between the organisations in an ecosystem may enable and improve active engagement of various stakeholders in the development of the central software technology.

When explaining the concept of software ecosystems it is also necessary to address how software ecosystems relate to the development of open source software [6]. There are clear similarities between these two concepts, but also several differences, which justify the definition of software ecosystems as a unique concept. The main difference between these two relates to the underlying business model. [3], explain the open-source business model as follows: *"The basic premise of an open-source approach is that by "giving away" part of the Company's intellectual property, you receive the benefits of access to a much larger Market. These users then become the source of additions and enhancements to the product to increase its value, and become the target for a range of revenue-generating products and services associated with the product."*

Whereas in a closed software ecosystem the intellectual property (the code) is *not* shared in any way.

However, different research directions indicated by literature and industrial cases re-enforce a lot of important perspectives to be explored, such as architecture, social networks, modelling, business, mobile platforms and organizational-based management [9]. Besides, SECOs involve a multidisciplinary perspective, including Sociology, Communication, Economy, Business and Law. These studies are also motivated by the software vendors' routine since they no longer function as independent units that can deliver separate products, but have become dependent on other software vendors for vital software components and infrastructures such as operating systems, libraries, component stores, and platforms [2].

VII. Architecture Of Major Software Ecosystems (SECOs)

1) Symbian Software Ecosystem

In this ecosystem as shown in figure 1, the different categories of licenses and partner relationships included are as shown:

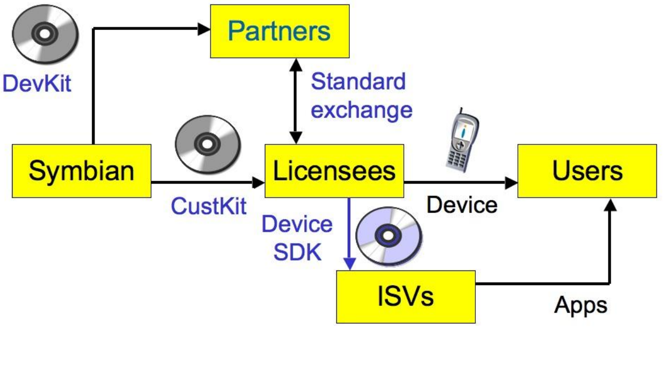


Fig. 1. Symbian Ecosystem [16]

Symbian described its network of customers and complementors as an “ecosystem”,

In the Symbian ecosystem, the different categories of licenses and partner relationships included are:

- System integrators or “licensees” (handset manufacturers) that integrated externally sourced software and internally developed hardware to create new devices (i.e. handsets) for sale to end users.
- CPU vendors worked to ensure Symbian OS compatibility with their latest processors.
- User Interface companies.
- Other software developers sometimes referred to as independent software vendors (ISVs) including developers of user applications and also middleware components such as databases.
- Network Operators, which in most countries were the dominant distribution channel for phones, and also decided what software components were preloaded on phones.
- Enterprise software developers, for cases where a company developed Symbian compatible software for its employees that use Symbian phones.

In many cases, members of Symbian’s ecosystem were also members of competing mobile phone ecosystems, such as those surrounding the Palm OS, Windows Mobile, and later Linux based platforms such as the LiMo Foundation and Google’s Open Handset Alliance (Android).

2) *Microsoft Software Ecosystem (SECO)*

Microsoft ecosystem consists of the following components: Device manufacturers, Independent Software Vendors (ISVs), Value Added Resellers (VARs), Office Equipment Dealers and Systems Integrators (SI) as shown in (Figure 2), and can all benefit from working together. But rarely do the ecosystem pieces remain static. New software applications are consistently being rolled out. And the VARs, dealers and SIs that sell and support these systems change with them.

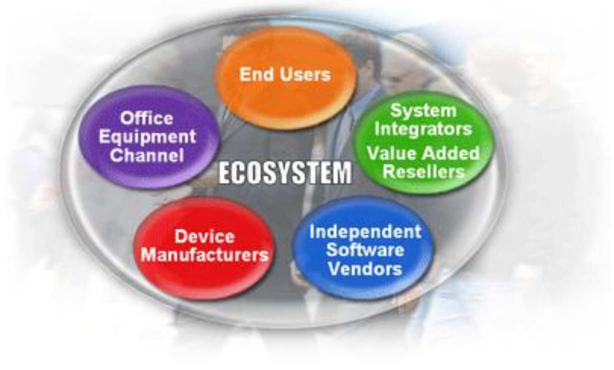


Fig. 2. Microsoft Software Ecosystem [7]

Microsoft sit at the centre of ecosystem. Ecosystems are an essential ingredient in delivering customer-focused solutions. And they help drive standards. And, they present revenue opportunities for all the partners involved. It’s no wonder that Microsoft spends so much money on building their ecosystem

The Microsoft ecosystem of applications, partners, and highly skilled IT resources provides customers with the best choice.

3) *iPhone Software Ecosystem*

The iPhone ecosystem which is one of the Apple’s three sub-ecosystems consists of the following components

- Developers and Designers
- Distribution
- Devices
- Users
- Internet
- Services and Advertisers

iPhone components are shown in figure3 below.

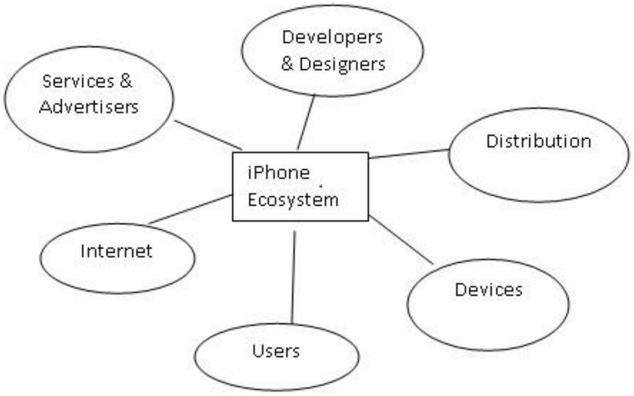


Fig. 3. iPhone components

Developers designs and implement complex interfaces smoothly and efficiently on limited hardware. C++ and

Objective-C are the primary languages used. Apple has historically put very little effort into supporting developers and designers, but has stepped up efforts for the iPhone platform. Designers are crucial to the success of iPhone applications. Developers simply utilise various technologies available to give designers what they want and need to build excellent interfaces.

4) *Ruby Software Ecosystem*

Ruby is a dynamic, open source programming language with a focus on simplicity and productivity. It has an elegant syntax that is natural to read and easy to write. It was created by Yukihiru Matsumota in 1995 in Japan.

The Ruby Software Ecosystem consists mainly of two elements i.e. Gems and Developers with possible relationships among them. If a developer has a relationship with a gem, he is a developer of that specific gem.

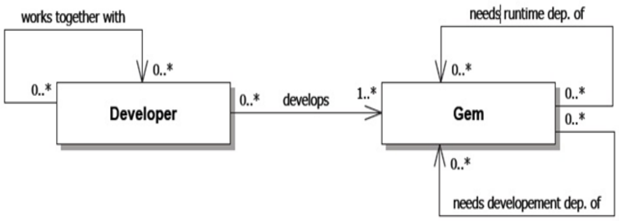


Fig. 4. Ruby Software Ecosystem [11]

The entire Ruby ecosystem consists of all developers, gems and their relationships as shown in figure 4. Some corporate high technology initiatives with Ruby are: **Sun Microsystems**, **Microsoft**, **Apple, IBM and SAP**.

5) Google Android Ecosystem

Android is a comprehensive open source platform designed for mobile devices. It is championed by Google and owned by Open Handset Alliance. The open Handset Alliance prominent members include: T-Mobile, Motorola, Samsung, Sonny Ericsson, Toshiba, Vodafone, Google, Intel, and Texas instrument. This list has grown multi fold with over 80 in number [5].

Android is revolutionizing the mobile space. It is a truly open platform that separates the hardware from the software that runs on it. This allows for a much larger number of devices to run the same applications and creates a much richer ecosystem for developers and consumers.

One way in which Android is quite different from other platforms is the distribution of its applications. On most other platforms, such as iPhone, a single vendor holds a monopoly over the distribution of applications. On Android, there are many different stores, or markets. Each market has its own set of policies with respect to what is allowed, how the revenue is split, and so on. As such, Android is much more of a free market space in which vendors compete for business. The figure 5 below summarised android software stack.

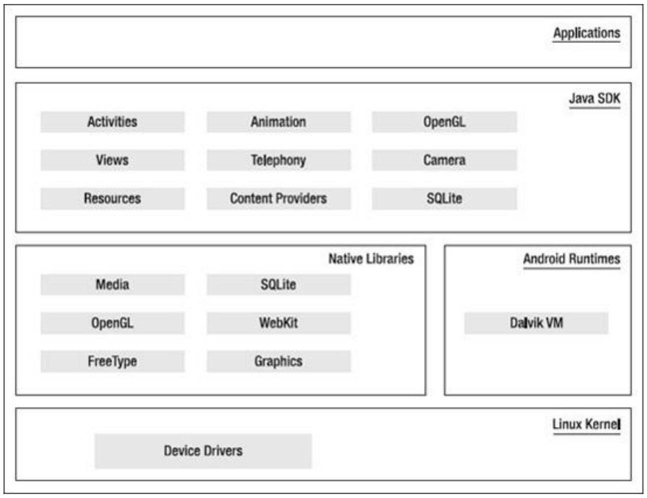


Fig. 5. Android Software Stack [13]

6) Eclipse Ecosystem

Eclipse is an open source integrated development environment (IDE) for Java. It was originally aimed to provide a united platform for different IDE products from IBM.

The Eclipse project, which began at the end of 1998, has an ambition to "eclipse" the leader of the IDE market. Within few years, Eclipse has evolved from Java IDE (version 1.0) to a universal tooling platform (version 2.0), and finally evolves to an application framework for building rich client application (version 3.0). Commercial software development tools such as IBM Rational tool, web sphere studio, and Borland JBuilder have been developed based on Eclipse.

Eclipse is currently managed by the Eclipse foundation with over 100 members including HP, IBM, Nokia, INTEL and Borland. The biggest challenge for the foundation is to cope with its rapid growth from its community.

Eclipse ecosystem Architecture

The functional building blocks of the Eclipse IDE are illustrated in Figure 6 below. The entire platform is open source and royalty-free for other open source or commercial products that add new building blocks.

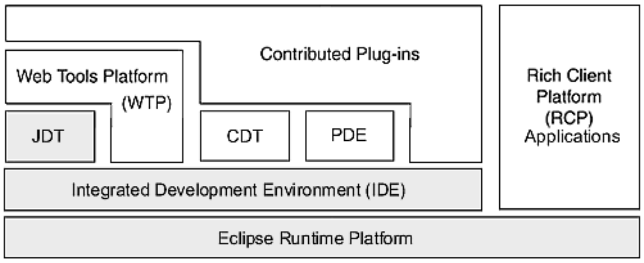

Contributed Plug-ins
Web Tools Platform (WTP)
JDT
CDT
PDE
Integrated Development Environment (IDE)
Rich Client Platform (RCP) Applications
Eclipse Runtime Platform

Fig. 6. Eclipse ecosystem Architecture [12]

A. Components of the Eclipse ecosystem Architecture

1. C/C++ Development Tools (CDT)

The C/C++ Development Tools (CDT) project is creating a fully functional C and C++ IDE for the Eclipse platform.

2. Plug-in Development Environment

The Plug-in Development Environment (PDE) supplies tools that automate the creation, manipulation, debugging, and deploying of plug-ins.

3. Java Development Tools

Java Development Tools (JDT) are the only programming language plug-ins included with the Eclipse SDK. However, other language tools are available or under development by Eclipse subprojects and plug-in contributors

4. Eclipse Runtime Platform

The core runtime platform provides the most basic level of services such as Loading plug-ins and managing a registry of available plug-ins, managing resources, update and help facility.

5. Integrated Development Environment

The Eclipse IDE provides a common user experience across multi-language and multi-role development activities.

6. Web Tools Platform

The mission of the Web Tools Platform (WTP) project is to provide a generic, extensible, and standards-based tool platform that builds on the Eclipse platform and other core Eclipse technologies.

7. Rich Client Platform

The Eclipse Rich Client Platform (RCP) is a set of plug-ins needed to build a rich client application.

The eclipse consortium is currently hosting eight top level projects and over thirty sub-level open source projects. There are also countless number of commercial and open source Eclipse related products, plug-ins, and distributions available from the internet. This virtual ecosystem takes care of software development, application life cycle, data management, and business operations

VIII. Open Source Software (Oss) And Closed Ecosystems - Similarities And Differences

TABLE II. The Similarities And Differences Between Open Source Software And Closed Systems

Similarities
A shared interest in the development, evolution, and use of a software product
Independent actors collaborate and contribute to development
Open innovation
New business models as compared to traditional licensed software

Differences

OSS	**Closed ecosystems**
Open source code.	Closed source code.
Ownership is shared.	Ownership and control lies with the keystone organisation.
Free use (with options for paying for specializations and related services)	Pay for use.
Extensibility through open source code.	Extensibility through controlled interfaces

IX. Features Of Software Ecosystems

The main features of SECOs are as follows.

1) They Inherits characteristics of natural ecosystems like mutualism, commensalism, symbiosis and so on

2) SECOs have architectural concepts like interface stability, evolution management, security and reliability

3) It is an to open source development model

4) They can be used to negotiate requirements for aligning needs with solutions, components, and portfolios

5) SECOs have capability for process innovation.

X. Benefits Of Software Ecosystems

1) Fosters the success of software co-evolution and innovation inside the organization involved and increases attractiveness for new players

2) Decreases costs involved in software development and distribution

3) Help analyse and understand software architecture

4) Supports cooperation and knowledge sharing among multiple and independent software vendors

5) Enables better analysis of requirements and communication among stakeholders

6) Help to overcome the challenges during design and maintenance of distributed applications

7) Provides help to the tasks of business identification, product architecture design and risk identification

8) Provides information for the product line manager regarding software dependencies

XI. Challenges Of Software Ecosystems

1) Establishing relationships between ecosystem actors and proposing an adequate representation of people and their knowledge in the ecosystem modelling.

2) Several key architectural challenges such as: platform interface stability, evolution, management, security, reliability.

3) Heterogeneity of software licenses and systems evolution in an ecosystem and how organizations must manage these issues in order to decrease risks of dependence.

4) Companies have difficulty at establishing a set of resources in order to differentiate from competitors.

5) Technical and socio-organizational barriers for coordination and communication of requirements in geographically distributed projects.

6) Insufficient infrastructures and tools for fostering social interaction, decision-making and development across organizations involved in both open source and proprietary ecosystems.

XII. Contributions

This paper contributes to the field of software ecosystems by providing

1) A necessary foundation for understanding how Software Ecosystems are composed and further aids understanding of this new and expanding area of software development.

2) A number of open research questions and challenges which should enable scholars interested in SECOs to swiftly gain an overview of this research area

XIII. Future Directions for Software Ecosystems

As with most novel approaches, this paper on SECO has opened up possibilities for new and exciting future directions. This following area should be investigated as future research directions/challenges for SECOs.

1) In Open source ecosystems.

a) How can quality be measured per developer?

b) How can relationships be formed between developers?

c) How can conflicts be resolved in open source ecosystems?

d) How can application program interfaces (APIs) to third-party components be used.

2) Governance.

a) What are the best strategies for survival in an ecosystem?

b) How can organisations involved achieve and maintain a healthy position in a SECO?

3) Analysis

a) How can an ecosystem be analysed.

b) Is it possible to create models, visualizations, and large data sets for analysis?

4) Openness

Every software platform at the centre of an ecosystem has to have some degree of openness. The main research question here is

How can openness in software affects and influences the success of a business, where there appears to be a real trade-off between the height of entry barriers and number of third parties willing to participate in the ecosystem.

5) Quality

a) How can ecosystems deliver the highest quality experience to customers in the ecosystem?

b) What are measures that participants can take to increase quality?

XIV. Conclusion

This paper provides a review of SECOs and confirmed that it is an emergent field that has been mainly inspired by studies from business and natural ecosystems. We highlighted that SECOs field needs more industrial studies to increase its body of evidence. Also, given the current state of research and practice in SECOs, we envisaged the need to conduct integrative studies among research communities and industry.

Finally the paper proposes a number of open research questions and challenges to enable scholars interested in SECOs to swiftly gain an overview of the research area and to help them in their own research endeavours.

References

[1] Bosch, J. (2009). From Software Product Lines to Software Ecosystems. In proceedings of 13th International Software Product Line Conference (SPLC'09), San Francisco, USA, 24-28 August.111-119.

[2] Boucharas, V., Jansen, S., and Brinkkemper, S., (2009), 'Formalizing Software Ecosystem Modeling'. In: Proceedings of the 1st International Workshop on Software Ecosystems, 11th International Conference on Software Reuse, Falls Church, USA, 34-48, September.

[3] Brown, A. W. and Booch, G. (2002). *Reusing Open-Source Software and Practices: The Impact of Open-Source on Commercial Vendors. In proceedings of 7th International Conference on Software Reuse:* Methods, Techniques, and Tools, Austin, USA, April 15-19. 123-136.

[4] Chesbrough, H. (2006). Open Innovation: A New Paradigm for Understanding Industrial Innovation. In Open Innovation: Researching a New Paradigm. Chesbrough, H., Vanhaverbeke, W. and West, J. (eds.). Oxford: Oxford University Press: 1-12.

[5] Fabio Cevasco (2011) Ruby Compendium: An essential Guide to the Ruby Ecosystem.

[6] Fitzgerald, B. (2006). The Transformation of Open Source Software. MIS Quarterly **30**(3): 587-598.

[7] Gantz J.F, Bibby D. (2011) White paper on Partner Opportunity in the Microsoft Ecosystem.

[8] Hanssen, G.K. and T.E. Fægri,(2008) Process Fusion -- Agile Product Line Engineering: an Industrial Case Study. Journal of Systems and Software **81**: p. 843---854

[9] Jansen, S., Brinkkemper S., Finkelstein A. Bosch J.(2009), Introduction to the Proceedings of the First Workshop on Software Ecosystems, in First International Workshop on Software Ecosystems. CEUR--WS.

[10] Jansen S., Brinkkemper S., Finkelstein, A.(2009) A Sense of community: A research agenda for software ecosystems. In: Proceedings of the 31st International Conference on Software Engineering.

[11] Kabbedijk, J., and Jansen, S., (2011), 'Steering Insight: An exploration of the Ruby Software Ecosystem'. In: Proceedings of the 2nd International Conference on Software Business, Brussels, Belgium, 44-55, June.

[12] Lam T., Gotz A. (2005)' Leveraging The Eclipse Ecosystem for Scientific Community'10th ICALEPCS Int. Conf. on Accelerator & Large Expt. Physics Control Systems. Geneva, 10 - 14 Oct 2005, TH3A.3-5O (2005)

[13] Mark Gargenta (2011) Learning Android: O'Reilly media Inc.

[14] Moore, J. F. (1993). Predators and prey: A new ecology of competition. Harvard Business Review 71(3): 75-86.

[15] Wirehead Labs, Inc. (2012). The iPhone Ecosystem

[16] Wood, David (2002). "Symbian Developer Expo 2002 - in context" internal presentation,Symbian Ltd., London.

[17] Xu, L., Brinkkemper, S. (2007): Concepts of product software. European Journal of Information systems 531-541

Improved QO-STBC OFDM System Using Null Interference Elimination

K. O. O. Anoh, R. A. Abd-Alhameed, Y. A. S. Dama, S. M. R. Jones, T. S. Ghazaany, J. Rodrigues and K. N. Voudouris
Mobile and Satellite Communications Research Centre, University of Bradford, UK BD7 1DP

***Abstract*—The quasi-orthogonal space time block coding (QO-STBC) over orthogonal frequency division multiplexing (OFDM) is investigated. Traditionally, QO-STBC does not achieve full diversity since the detection matrix of QO-STBC scheme is not a diagonal matrix. In STBC, the decoding matrix is a diagonal matrix which enables linear decoding whereas the decoding matrix in traditional QO-STBC does not enable linear decoding. In this paper it is shown that there are some interfering terms in terms of non-diagonal elements that result from the decoding process which limit the linear decoding. As a result, interference from the application of the QO-STBC decoding matrix depletes the performance of the scheme such that full diversity is not attained. A method of eliminating this interference in QO-STBC is investigated by nulling the interfering terms towards full diversity for an OFDM system. It was found that the interference reduction technique permits circa 2dB BER performance gain in QO-STBC. The theoretical and simulation results are presented, for both traditional QO-STBC and interference-free QO-STBC applying OFDM.**

***Keywords*—QO-STBC; STBC; OFDM; Decoding matrix; Null-Interference; Interference;**

I. Introduction

Present days wireless communications systems are in great quest for efficient communications. Wi-Fi and terrestrial base stations are increasingly deploying the multi-antenna system for seamless communications. For instance, the multiple input multiple output (MIMO) antenna configuration is useful in achieving higher throughput in these wireless communication systems. Space Time block coding (STBC) is one of interesting methods for deploying this technique. The advantage of using, for example, the orthogonal STBC (OSTBC) is that it exploits full power transmission for orthogonal codes so long as the transmitter diversity order is no more than two [1-3]. For more than two transmit diversity, it has been shown that full rate power is not possible [4]. Meanwhile, it is possible to deploy the STBC technology in way that full rate power transmission can be achieved. In such case, the codes are rather formed in a special orthogonal way. This is usually discussed as the quasi-orthogonal STBC, hereinafter QO-STBC. The QO-STBC offers the advantage of improved channel capacity and also improved bit error ratio (BER) statistics for a multi-antenna transmission [2].

In [1, 5, 6], a QO-STBC was introduced. It achieves full transmission rate but not full diversity [7]. QO-STBC sacrifices both BER measure with increasing SNR and full transmission diversity but offers excellent transmission rate. The BER curves suggest that the codes outperform the codes of orthogonal design only at low SNRs, but worsen at increased SNRs [1]. This is due to the fact that the slope of the performance curve depends on the diversity order gain. One of the major problems that limits the BER performance of the QO-STBC system is from the interference incurred in the decoding process. If these interferences can be removed, then the performance of the QO-STBC scheme will improve towards full diversity gain.

An example of the approach deployed towards achieving full diversity by interference reduction has been shown in [2, 8]. The method involves nulling the interfering terms in the resulting decoding process to improve the performance of the scheme. Then, it is well known that combining OFDM with MIMO thrives towards achieving improved BER and better throughput [9]. The OFDM treats frequency selective channel as flat fading channel with cyclic prefix (CP) that is at least equal to the channel delay spread. This is used to overcome inter-symbol interference (ISI). In this study, the channel model is limited to a correlated multipath transmission only. In such case, the design exploits multipath channel gains due to multipath of flat fading transmission or frequency non-selective channel. The individual path gains for a particular transmission branch are assumed to be uniform.

In this study, the QO-STBC method is studied for three and four transmit antennas. Using the interference elimination approach, it will be shown that the behaviour of QO-STBC can be improved.

The OFDM system and channel model are discussed in Section II and the full transmission rate QO-STBC is discussed in Section III. In Section IV, the free interference QO-STBC with full diversity while OFDM based QO-STBC architecture is presented in Section V. The numerical simulation results for QO-STBC OFDM system is discussed in Section VI followed by summarized conclusion.

II. The OFDM System And Channel Model

OFDM divides wideband into many narrow-bands and treat the channel as a flat fading channel. In this section, both the OFDM scheme and the channel model are presented.

A. *Orthogonal Frequency Division Multiplexing (OFDM) Technology*

In the baseband of multicarrier system, OFDM multiplexing scheme is used to divide a selected wideband spectrum into many smaller narrow bands. This is achieved using the fast Fourier transform (FFT). Over a frequency selective channel, a predefined length of the symbol is used to overcome inter-symbol interference (ISI) for channel impulses with long delay. This symbol length, usually called the cyclic prefix (CP) is usually longer (or least equal to) the length of the worst delay in time. In time domain, an *N*-point FFT OFDM system can be defined as:

$$b[n] = \frac{1}{\sqrt{N}} \sum_{k=0}^{N-1} B_{(k,l)} e^{j2\pi nk/N}, \quad n = 0, 1, 2, \ldots, N-1 \qquad (1)$$

In Equation 1, *N* is the number of narrowband sub-channels, $1/\sqrt{N}$ is a scaling factor with *n* as the index of the prevalent subcarrier. $B_{(k,l)}$ is the kth sub-channel input symbol of the lth - constellation mapped using, for instance, QPSK. Meanwhile, the number of FFT points adopted in the design of any specific OFDM structure provides the number of narrowband sub-channels over which the input symbols are multiplexed. Since the design of QO-STBC requires at least 4 constellations, then the QPSK is applied in this study. Considering the CP, Equation 1 is modified to include the guard length as;

$$b[n] = \sqrt{\frac{N}{N+N_g}} \sum_{k=0}^{N-1} B_{(k,l)} e^{j2\pi nk/N}, -N_g \le n < N \qquad (2)$$

Where N_g is the pre-appended cyclic prefix length. This CP is used to combat inter-symbol interference (ISI) which is caused by multipath delay.

B. Channel Model

The channel model of QO-STBC involves N_T-transmit and N_R-receive antennas. The channel is a typical correlated (non-frequency selective fading) multipath channel with *L*-independent propagation paths and the same power-delay profile. If $h_{i,k}(n)$ represent the channel impulse response, there will be a vector of $[h_{i,k}(0)\ h_{i,k}(1)\ h_{i,k}(2)\ldots h_{i,k}(L\text{-}1)] \in C^{1\times L}$ independent channel taps corresponding to ith-receive antenna and kth-transmit antenna, where $C^{1\times L}$ is a complex vector. Each of the taps is generally represented according to [10] thus;

$$h_{i,k}(n) = \sum_{l=0}^{L-1} \alpha_{i,k}(n)\, \delta(t-\tau_l)\, e^{j\theta_l} \qquad (3)$$

Where α is the gain, $\theta = (2\pi l n/N)$ is the phase and τ the path delay corresponding to lth-path. $0 \le n \le N\text{-}1$ characterizes the independent OFDM frequency subcarriers. The channel impulse response of Equation 3 for each path is a zero mean complex Gaussian random variable with a normalized variance of unity. For QO-STBC, the codewords are modulated independently for each transmission branch. For instance, if there are four different time slot codewords then there must be four independent OFDM modulators, one for each. For maximum diversity gains, it is required that number of OFDM subcarriers be more than or equal to the number of independent delay paths, *L* [11]. Now, if ***C*** is the independent codeword resulting in the discrete $b_j[n]$-term after OFDM modulation, then the received message will be:

$$r^t_k(n) = \sum_{i=1}^{N_T} b^t_j(n) H^t_{ik}(n) + Z_k(n), \qquad 0 \le n \le N \qquad (4)$$

Notice that $H^t_{ik}(n)$ is the channel transfer function (frequency response) which can be represented as;

$$H^t_{ik}(n) = \sum_{l=0}^{L-1} h^t_{i,k}(n) e^{-j2\pi l n/N}, \quad j = \sqrt{-1},\ 0 \le n \le N \qquad (5)$$

In Equation 4, $Z_k(n)$ is the Gaussian noise term of the nth-OFDM subcarrier at tth OFDM duration. The channel model is assumed to be frequency non-selective across all frequencies.

III. Conventional QO-STBC With Full Rate

The well-known conventional QO-STBC with full transmission rate for three antennas or more [1, 6] encode matrix columns by dividing them into two orthogonal groups. The codes in the columns within each group are not orthogonal but the codes from columns of different groups are orthogonal. This scheme does not achieve full diversity due to some coupling terms between the estimated symbols [8]. The coupling terms are the problems that deplete the BER performance of the QO-OSTBC scheme. However, using the interference mitigation technique reported in [2, 8], improved BER performance can be achieved without losing the full rate transmission and diversity. Meanwhile, let the conventional QO-STBC defined in [6] be expressed as:

$$C = \begin{bmatrix} \Omega_{12} & \Omega_{34} \\ \Omega^*_{34} & \Omega^*_{12} \end{bmatrix} = \begin{bmatrix} c_1 & c_2 & c_3 & c_4 \\ -c^*_2 & c^*_1 & -c^*_4 & c^*_3 \\ c_3 & c_4 & c_1 & c_2 \\ -c^*_4 & c^*_3 & -c^*_2 & c^*_1 \end{bmatrix} \qquad (6)$$

where Ω represents the traditional Alamouti orthogonal space time block codes [4] defined as:

$$\Omega_{12} = \begin{bmatrix} c_1 & c_2 \\ -c^*_2 & c^*_1 \end{bmatrix} \text{and } \Omega_{34} = \begin{bmatrix} c_3 & c_4 \\ -c^*_4 & c^*_3 \end{bmatrix} \qquad (7)$$

Equation 6 constructs a code of *P* constellations transmitted at *T* time slots. Here, there are 4 constellations and 4 time slots so that the achieved transmission rate *R* = 1 (full rate) where *R* = *P*/*T*. It must be emphasized that the entries [c_1, c_2, c_3, c_4] of the matrix ***C*** are related to *b*-terms by the OFDM transformation and should not be confused. However, the QPSK mapping scheme is applied in this study.

Following the tradition in [3, 12], the equivalent virtual channel matrix following Equation 6 can be defined as:

$$H_v = \begin{bmatrix} h_1 & h_2 & h_3 & h_4 \\ h^*_2 & -h^*_1 & h^*_4 & -h^*_3 \\ h_3 & h_4 & h_1 & h_2 \\ h^*_4 & -h^*_3 & h^*_2 & -h^*_1 \end{bmatrix} \qquad (8)$$

Let there be N_R maximum receive antennas and N_T maximum transmit antennas (It is assumed N_T = 4 in this study). Then, the received signal at t time slot and on i-receive antenna ($1 \leq i \leq N_R$) can be expressed in a simplified form as:

$$r_{t,i} = \sum_{l=0}^{L-1} \alpha_{l,i}\, b_{t,l} + Z_{t,i} \qquad (9)$$

$\alpha_{n,l}$ is the path gain of the l^{th} path and $Z_{t,i}$ is the additive white Gaussian noise (AWGN) and necessarily a matrix (vector) of the form:

$$Z_{t,i} = \begin{bmatrix} z_1 \\ -z_2^* \\ z_2 \\ z_1^* \end{bmatrix} \qquad (10)$$

The path gains are contributed from the impulse responses of the channel individually obtained as in Equation 3. Notice that $h_{i,k}(n)$ belongs to the entries of the virtual channel matrix, H_v of Equation 8. This study is a case of a flat-fading transmission with the impulse responses of the channel individual path being correlated, hence,

$$r_{t,l} = H_v\, b + Z \qquad (11)$$

Equation 11 represents the received symbols at t^{th} -time slot on i^{th} – receive antenna of the receiver. Now, let the decoding method proceed as:

$$\hat{b} = H_v^H \cdot r = H_v^H H_v \cdot b + H_v^H \cdot Z \qquad (12)$$

Or,

$$\hat{b} = D_4 \cdot b + H_v^H \cdot Z$$

Where D_4 is the quasi-orthogonal detection matrix for four transmit antennas and H_v^H is the Hermitian equivalent of the matrix H_v. This conventional quasi-orthogonal detection matrix is defined as:

$$D_4 = \begin{bmatrix} \lambda & 0 & \beta & 0 \\ 0 & \lambda & 0 & \beta \\ \beta & 0 & \lambda & 0 \\ 0 & \beta & 0 & \lambda \end{bmatrix} \qquad (13)$$

Equation 13 represents the estimate of the codes that are inseparable, one from another, and the interference terms also. In orthogonal STBC, the detection matrix, D, is always a diagonal matrix, and so enables a simple linear decoding [13]. This is not possible in QO-STBC since the resulting detection matrix (for instance, Equation 13) is not orthogonal and so not a diagonal matrix, instead quasi-orthogonal. Therefore the simple linear decoding cannot be implemented. Thus, an interference free quasi-orthogonal detection matrix must be of the form [2, 8]:

$$D = \begin{bmatrix} \lambda+\beta & 0 & 0 & 0 \\ 0 & \lambda+\beta & 0 & 0 \\ 0 & 0 & \lambda-\beta & 0 \\ 0 & 0 & 0 & \lambda-\beta \end{bmatrix}$$

Where λ is the diagonal of the (4 × 4) I_4 matrix. This is the sum of the channel power (or the path gains) and represented as $\lambda = \sum_{n=1}^{N} \left\| h_n \right\|^2, \quad \forall n = 0,1,2,3,4$. Also, β represents the interfering terms that deplete the full diversity performance expected of the 4-transmit antenna elements and is computed as: $\beta = h_1 h_3^* + h_2 h_4^* + h_1^* h_3 + h_2^* h_4$.

Thus, β will degrade the BER performance of the system as long as the aforementioned decoding approach is followed. This can be improved by using a more complex decoding approach such that the better estimate of the transmit symbol be obtained [8, 13]. Let the effective estimate of the transmit symbols be:

$$\begin{aligned} \breve{b} &= (H_v^H \cdot H_v)^{-1} H_v^H \cdot r \\ &= (H_v^H \cdot H_v)^{-1} D_4 \cdot b + (H_v^H \cdot H_v)^{-1} H_v^H \cdot Z \\ &= (H_v^H \cdot H_v)^{-1} H_v^H H_v \cdot b + (H_v^H \cdot H_v)^{-1} H_v^H \cdot Z \end{aligned} \qquad (14a)$$

So that,

$$\breve{b} = b + (H_v^H \cdot H_v)^{-1} H_v^H \cdot Z \qquad (14b)$$

$\breve{b}$ is the effective received symbol after channel compensation. This symbol contains the interfering terms which deplete the BER performance. In the next section, a method of eliminating the interference term discussed in [2, 8] is discussed.

IV. Free-Interference QO-Stbc With Full Diversity

In this section, the approach for reducing the coupling (interfering) terms in the decoding matrix of the traditional QO-STBC system is discussed. The code structure adopted has been discussed in [8] following the QO-STBC method of code construction discussed in [6]. An interference free QO-STBC system achieves full diversity (see [11]) and the decoding matrix should be of the Equation 15 form [2, 13]. Equation 14 can be obtained as in [2, 8] by formulating:

$$D = \begin{bmatrix} \lambda+\beta & 0 & 0 & 0 \\ 0 & \lambda+\beta & 0 & 0 \\ 0 & 0 & \lambda-\beta & 0 \\ 0 & 0 & 0 & \lambda-\beta \end{bmatrix} \qquad (15)$$

This can be obtained as in [2, 8] by formulating:

$$D_4 \cdot V - V \cdot D = 0 \qquad (16)$$

Where V is the eigenvector as:

$$V=\begin{bmatrix}1 & 0 & -1 & 0\\ 0 & 1 & 0 & -1\\ 1 & 0 & 1 & 0\\ 0 & 1 & 0 & 1\end{bmatrix} \tag{17}$$

From Equation 16,

$$\begin{aligned} D &= V^{-1}D_4V\\ D &= \frac{1}{2}V^H\cdot D_4\cdot V\\ D &= \frac{1}{2}V^H\cdot H_v^H H_v\cdot V \end{aligned} \tag{18}$$

Where $V^{-1}=\frac{1}{2}V^H$.

Now the resulting virtual channel matrix would be;

$$H_{new}=H_v\cdot V$$

$$=\begin{bmatrix} h_1+h_3 & h_2+h_4 & h_3-h_1 & h_4-h_2\\ h_2^*+h_4^* & -h_1^*-h_3^* & h_4^*-h_2^* & h_1^*-h_3^*\\ h_1+h_3 & h_2+h_4 & h_1-h_3 & h_2-h_4\\ h_2^*+h_4^* & -h_1^*-h_3^* & h_2^*-h_4^* & h_3^*-h_1^* \end{bmatrix} \tag{19}$$

Following the channel matrix of Equation 17 and the tradition in [3, 12], the encoding matrix can be formulated as:

$$C_{new}=\begin{bmatrix} c_1-c_3 & c_2-c_4 & c_3-c_1 & c_4+c_2\\ c_4^*-c_2^* & -c_3^*+c_1^* & -c_4^*-c_2^* & c_3^*+c_1^*\\ c_1-c_3 & c_4+c_2 & c_3-c_1 & c_4-c_2\\ -c_4^*-c_2^* & c_3^*+c_1^* & c_4^*-c_2^* & -c_3^*+c_1^* \end{bmatrix} \tag{20}$$

By deleting the fourth column of the new matrix, the new matrix for a 3x1 QO-STBC can be formed thus;

$$\begin{bmatrix} c_1-c_3 & c_2-c_4 & c_3+c_1\\ c_4^*-c_2^* & -c_3^*+c_1^* & -c_4^*-c_2^*\\ c_1+c_3 & c_2+c_4 & c_3-c_1\\ -c_4^*-c_2^* & -c_3^*+c_1^* & c_4^*-c_2^* \end{bmatrix} \tag{21}$$

By nulling the fourth antenna elements of the channel matrix in Equation 8, then the equivalent channel matrix for 3×1 can be formed.

$$H_{new3}=\begin{bmatrix} h_1+h_3 & h_2 & h_3-h_1 & -h_2\\ h_2^* & -h_1^*-h_3^* & -h_2^* & h_1^*-h_3^*\\ h_1+h_3 & h_2 & h_1-h_3 & h_2\\ h_2^* & -h_1^*-h_3^* & h_2^* & h_3^*-h_1^* \end{bmatrix} \tag{22}$$

As it would be expected, the 4 x 1 QO-STBC outperforms the 3 x 1QO-STBC as shown in Figure 1.

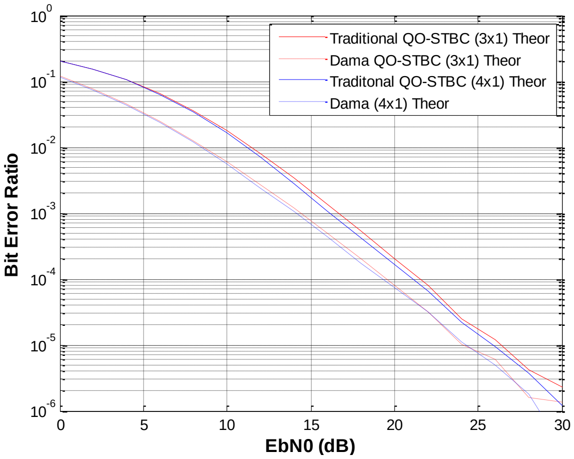


Fig 1. Comparison of Traditional QO-STBC and Interference-free QO-STBC

It is observed that interference elimination approach achieves circa 2dB gain with respect to the traditional QO-STBC for both 4 × 1 QO-STBC and 3 × 1 QO-STBC. This approach is then extended to include the OFDM as discussed in the next section.

V. The QO-Stbc Ofdm System Model

Combining the QO-STBC scheme and the OFDM system presented in Section II, the QO-STBC OFDM architecture is represented in Figure 2. The combination of QO-STBC scheme with OFDM drives the system towards achieving maximum diversity gain since in achieving maximum diversity gain, the number of subcarriers, *N*, must be larger than or equal to the number of independent delay paths, *L*, [11].

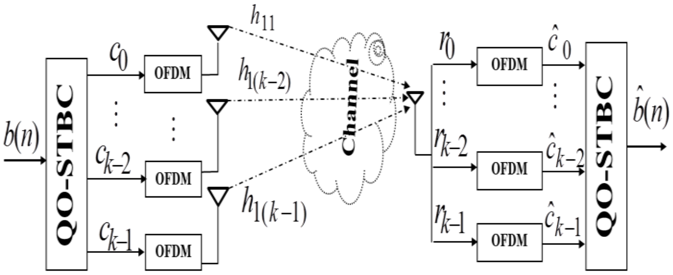


Fig 2. Design Architecture for QO-STBC OFDM System

This architecture is shown for a QPSK system with *b*(*n*) as the resulting QPSK mapped symbols in the transmitter. Its encoding is according to the descriptions of the QO-STBC above. According to the number of the required transmit antennas, the output QO-STBC coded symbols C_0, …, $C_{k\text{-}2}$, $C_{k\text{-}1}$ (where *k* is the maximum number of transmit antennas) are individually modulated by the OFDM modulation scheme. After traversing the channel the OFDM symbols are received as $r_0, \cdots, r_{k-2}, r_{k-1}$ corresponding to each transmission branch. Then, the estimates are demodulated by the OFDM of each transmission branch to obtain the estimates of the QO-STBC encoded symbols, $\hat{c}_0, \cdots, \hat{c}_{k-2}, \hat{c}_{k-1}$, for onward QO-STBC decoding. These estimates of the OFDM modulated symbols are obtained by performing an *N*-point inverse Fourier transform (IFFT) and removing the CP. After decoding, an estimate of the transmitted symbol, $\hat{b}(n)$, is obtained. This is fed to the QPSK de-mapping block and then for error computation.

VI. Numerical Simulation Results And Discussion

In the simulation, the symbols are constructed according to the codes. The codes provide, for three transmit antenna system, three time slots. So, the QO-STBC OFDM system symbols will be quasi-static for three slots. Also, for four time slots where no column of the encoding matrix is eliminated, then the QO-STBC OFDM symbols will be quasi-static for four time slots. The OFDM was designed with a CP length of 25%. Recall that one of the major advantages of OFDM system is that it converts a frequency selective transmission channel to a non-frequency selective channel such that correlation difference of one channel impulse response to another is highly negligible. In addition to this advantage, this study investigated a scenario where the channel is necessarily correlated. This condition is described for a frequency non-selective transmission. Thus, the flat fading channel has the characteristics that the channel coefficients are strictly uniform for each transmission branch. The results in Figure 3 are identical in behaviour to the results earlier reported in Figure 1. It is shown that for all cases the QPSK QO-STBC interference-free OFDM systems outperformed the traditional QPSK QO-STBC OFDM systems. Notice that the interference free method is designated in Figure 3 after the name of the proposer, Dama et al.

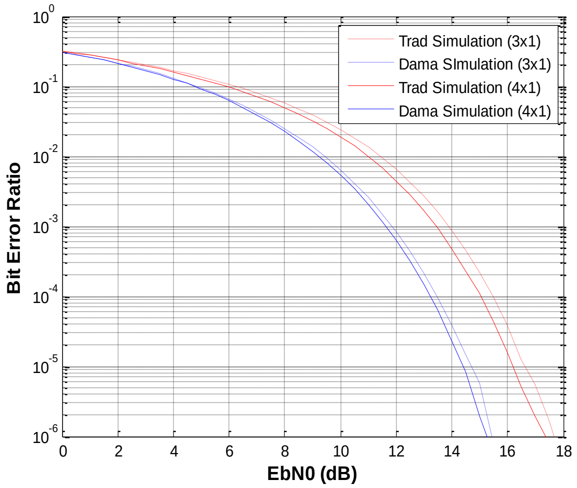


Fig 3. Comparison of Simulated results of interference QO-STBC OFDM system with traditional QO-STBC OFDM system

It is necessary to emphasize that the OFDM is characterized with better performance than the theoretical result. In QO-STBC OFDM, diversity advantage improves with the number of subcarrier. Also, the presence of cyclic prefix provides a phenomenon that reduces the irreducible error that is not possible in the traditional systems. In that sense, the inter-symbol interference are well overcome.

VII. Conclusion

QO-STBC scheme that improve the performance of multi antenna system using diversity over OFDM system has been presented. Traditional QO-STBC systems achieve only full transmission rate but not full diversity. The full diversity limitation has been shown to be consequent on the interfering terms incurred during decoding. These interfering terms exist as off-diagonal elements in the decoding matrix, which when there is no interference the matrix is a diagonal matrix and so permit linear decoding. An interference elimination technique was used in the case of a QO-STBC OFDM system to improve the performance of the QO-STBC scheme over an OFDM system. The decoding matrix of the QO-STBC is constructed for a free-interference decoding which improves the system performance. It was observed that the scheme achieved about 2dB gain both in theory and when OFDM system was applied during simulation. Consequently, the interference elimination method must be considered in the deployment of QO-STBC for multi antenna transmission in OFDM systems.

Acknowledgment

This work was partially sponsored by Datong plc, UK as part of an on-going research project and to be employed as part of their future communication services and financially by the Ebonyi State Government of Nigeria. The authors wish to thank the Ebonyi State Government for their support.

References

[1] H. Jafarkhani, "A quasi-orthogonal space-time block code," IEEE Transactions on Communications, vol. 49, pp. 1-4, 2001.

[2] Y. A. S. Dama, R. A. Abd-Alhameed, S. M. R. Jones, H. S. O. Migdadi, and P. S. Excell, "A NEW APPROACH TO QUASI-ORTHOGONAL SPACE-TIME BLOCK CODING APPLIED TO QUADRUPLE MIMO TRANSMIT ANTENNAS," Fourth International Conference on Internet Technologies & Applications, Sept. 2011, 2011.

[3] A. Goldsmith, Wireless communications: Cambridge university press, 2005.

[4] S. M. Alamouti, "A simple transmit diversity technique for wireless communications," Selected Areas in IEEE Journal on Communications, vol. 16, pp. 1451-1458, 1998.

[5] C. B. Papadias and G. J. Foschini, "Capacity-approaching space-time codes for systems employing four transmitter antennas," IEEE Transactions on Information Theory, vol. 49, pp. 726-732, 2003.

[6] O. Tirkkonen, A. Boariu, and A. Hottinen, "Minimal non-orthogonality rate 1 space-time block code for 3+ Tx antennas," in 2000 IEEE Sixth International Symposium on Spread Spectrum Techniques and Applications, 2000, pp. 429-432.

[7] W. Su and X.-G. Xia, "Signal constellations for quasi-orthogonal space-time block codes with full diversity," IEEE Transactions on Information Theory, vol. 50, pp. 2331-2347, 2004.

[8] Y. Dama, R. Abd-Alhameed, T. Ghazaany, and S. Zhu, "A New Approach for OSTBC and QOSTBC," International Journal of Computer Applications, vol. 67, pp. 45-48, 2013.

[9] K. O. O. Anoh, R. A. Abd-alhameed, J. M. Noras, and S. M. R. Jones, "Wavelet Packet Transform Modulation for Multiple Input Multiple Output Applications," IJCA, vol. 63 - Number 7, pp. 46 - 51, 2013.

[10] D. Tse and P. Viswanath, Fundamentals of wireless communication: Cambridge university press, 2005.

[11] F. Fazel and H. Jafarkhani, "Quasi-orthogonal space-frequency and space-time-frequency block codes for MIMO OFDM channels," IEEE Transactions on Wireless Communications, vol. 7, pp. 184-192, 2008.

[12] J. Proakis and M. Salehi, Digital Communications, Fifth ed. Asia: McGraw-Hill, 2008.

[13] U. Park, S. Kim, K. Lim, and J. Li, "A novel QO-STBC scheme with linear decoding for three and four transmit antennas," IEEE Communications Letters, vol. 12, pp. 868-870, 2008.

Pilot Study of Industry Perspective on Requirement Engineering Education: Measurement of Rasch Analysis

NOR AZLIANA AKMAL JAMALUDIN
Faculty of Computer Science and Information System
Universiti Selangor
45600 Bestari Jaya, Selangor
MALAYSIA

SHAMSUL SAHIBUDDIN
Advanced Informatics School (AIS), UTM International
Campus, Universiti Teknologi Malaysia (UTM)
Jalan Semarak, 54100 Kuala Lumpur
MALAYSIA

***Abstract*—Software development industry identifies that human-based give a significant problem in Requirement Engineering. To that reason, education gives a substantial impact in delivering a skill worker and should be a medium to reduce the problem. Survey question was distributed among ICT for this pilot study to the organization of MSC status in Malaysia for pilot study. 15.53% (N = 32) respondent successfully return their respond back. The result shows that only 27 person is analyzed regarding to misfit data provided by Rasch Measurement Model. The unidimensionality, person-item map and misfit data are discussed. Research objective to identify the undergraduate problem in Requirement Engineering education is achieved. Future work will be discussed on further analysis on actual survey to improve employability skill among software engineering undergraduate students.**

Keywords—Higher Learning Education; Requirement Engineering; Education; Rasch Measurement Model; Employability Skill; Undergraduate Problem; Rasch Analysis; Unidimensionality; Human-based Problem.

I. INTRODUCTION

Requirement Engineering is a fundamental process to meet stakeholder's needs in software development project. Failure in meeting stakeholder satisfaction will contribute to delay software development project, waste of time, energy, resources and poor quality [22]. The argument of this matter was widely discussed relying on the industry perspective. Industry was spent lots of money in research and development to identify the problem occurs in establishing a strong rapport within stakeholder.

Human-based give a weighty problem to the Requirement engineering. The classification of requirements problems include 1) lack of customer, user and developer 2) lack of communication 3) lack of training 4) lack of define responsibility 5) unstable workforce (low staff retention) 6) inappropriate skills 7) poor time and resource allocations [22].

To that reason, education should be a medium to reduce the human-based problem.

II. BACKGROUND

Requirement engineering education can be a weapon to sharpen the human-based skill particularly. Five evidence shows that Prophet Muhammad S.A.W. is an educator that takes into account individual differences in delivering teaching that consist of [1]: 1) provide appropriate advice in accordance with the difference that each seeking some advice 2) give a different answer to the same question that tailored to the individual who asked 3) behave and be different accordingly to the suitability of the mix therewith 4) deliver and legal adapted to the ability for the person to receive it and 5) implement and receive behavior but a person nor receive from someone else because of the different situations.

Each Higher Learning Education (HLE) has set objectives differently. Towards CGPA or student result, it goes the same meaning for all HLE. Cumulative Grade Point Average or CGPA is used widely in a developing country. Many people assuming that the highest CGPA student will have the highest performance to show their competency [9, 10, 11].

However, the validity of Cumulative Grade Point Average (CGPA) is purely the mean of raw scores, lack precision and linearity [8] to meet criteria for measure human-based skill should be revised back. The objective of this paper is to identify the problem in the current Requirement Engineering education practice using Rasch Measurement Model.

Figure 1 depicts the current practice process in Higher Learning Education (HLE) that give an input to the current problem arose in HLE. This practice will further analyze in Figure 7.

III. METHODOLOGY

The survey was administered randomly from 2201 ICT Malaysia Status Company (MSC) that registered under Multimedia Development Corporation (MDeC). The companies categorized from Shared Service Outsourcing, InfoTech, Creative Multimedia and IHL & Incubator industry [MDEC, 2012]. Klang Valley is chosen as a location for this pilot study because of an easy accessible respondent and majority MSC organization is located at Klang Valley.

The main objective of this pilot study is to identify the correctness of the questionnaire. Somehow, the questionnaire is expected to give a similar meaning (consistency) to all respondent. If not, some of the questionnaire will be rephrased or removed based on the analysis after data collection in this pilot study. To that reason, consistency in response will be analyzed. In addition, pilot test is used to know an expected outcome based on difficult and easy task respond from the respondent.

All observations begin as counts. But raw counts are only indications of a possible measure. Raw counts cannot be the measures sought because in their raw state, they have little inferential value. To develop metric meaning, the counts must be incorporated into a stochastic process which constructs inferential stability.

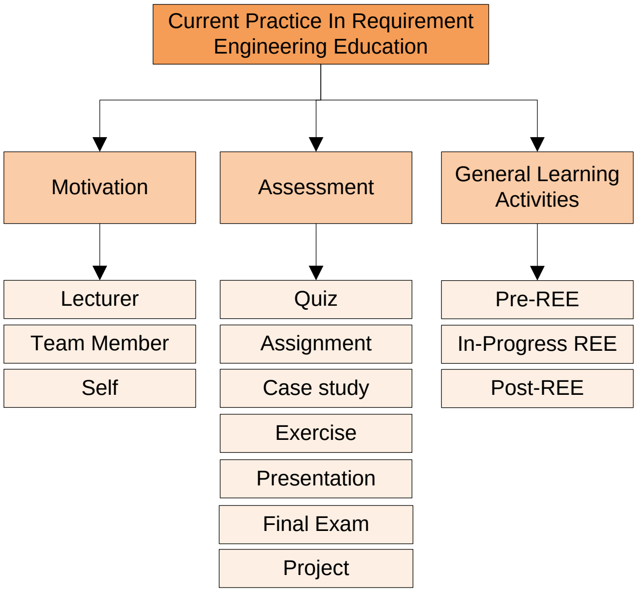


Fig. 1. Process of Current Practice in HLE

Thus, in order to construct inference from observation, the measurement model must: (a) produce linear measures, (b) overcome missing data, (c) give estimates of precision, (d) have devices for detecting misfit, and (e) the parameters of the object being measured and of the measurement instrument must be separable. Only the Rasch measurement models solve these problems. Rasch provide empirical evidence.

IV. Validity of Instrument

The questionnaire is revision based on the Cognitive Domain [4], Bloom’s Revised Taxanomy [17] for undergraduate practice [10]. Content [7,12], construct [3] and predictive [13] validity is crucial in getting the sufficient result of the study. Survey questionnaire in this research used two types of scale. First, the dichotomous scale [18] that comprise of 1 – 2 scale for ‘Yes’ and ‘No’. The scale is based on management style of decision making. Second, Likert [23] scale (1 – 2 – 3 – 4 – 5) used for descriptive response categories (never - rarely - sometime - often - always) as a means of partitioning the underlying latent quantitative continuum into successively increasing (or decreasing) amounts of the variable [2].

It leads wrong contextualize the result if we only using ordinal data to achieve the sufficient result. As a solution, a reliable instrument is needed. Rasch shows differently. Rasch transform an ordinal data which is qualitative data into ratio data. Rasch Analysis is deployed vigorously. It used to achieve an effective instrument construct of precision. Rasch able to sieve the instrument clean from any item misfit hence potential data defects [19, 20].

V. Rasch Measurement Model

Rasch Model is used to analyze data. Application of the Rasch model through software such as Winstep [14] and other Rasch software provide estimates of person and threshold locations on the latent variable scale. The software also yields indices of item and person fit to show that the requirement of unidimensionality is met.

Rasch answer on how to have the right measurement with valid instrument. Instrument is extremely crucial if involve human life. Based on Linacre (2011), things would change appreciably when you want a thermometer fit for an open heart surgery. Certainly we will need a more precise measurement instrument. Life is at stake, so it is necessary to have the correct instrument in place. Cost is no more the issue; precision and reliability overrides all. So it goes in an instrument construct; the Standard Error of Measurement and Item Reliability matters most that ought to be given priority when it comes to high stake measurement.

A. *Rasch Analysis*

The normal solution is to apply the regression approach. It shows the best fit line that inline with the points as best as possible. Then, it can be used to make the required predictions by interpolation or extrapolation [8] as necessary as shown in Figure 2.

$$y = \beta 0 + \beta 1 m \qquad \text{Equ .. (l)}$$

In obtaining the best fit line, there exist differences between the actual point; y and the best line, the predicted point; ý. The difference is referred to as error; e.

$$yi – ýi = ei \qquad \text{Equ .. (2)}$$

By accepting the fact that there is always error involved in the prediction model, the deterministic model of equation: 1) can be transformed into probabilistic model by including the prediction error into the equation; Equ. 3) Rasch moves the concept of reliability from establishing "best fit line" of the data into producing reliable repeatable measurement instrument. Rasch focuses on constructing the measurement instrument rather than fitting the data to suit the measurement model. $y = \beta 0 + \beta 1 m + e$ Equ .. (3)

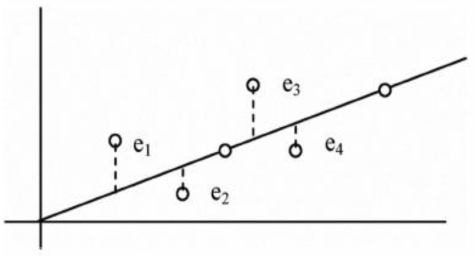


Fig. 2. Best fit line: Linear Regression Model

VI. Discussion

Only 15.53% (N = 32) personnel were successfully returned the survey question. Majority feedback is coming from Shared Service Outsourcing. The services sector is a vital contributor to the growth of the Malaysian economy and functions include Information Technology (IT) services, shared services and business process outsourcing (BPO), regional headquarters, research and development (R&D), training and environmental management.

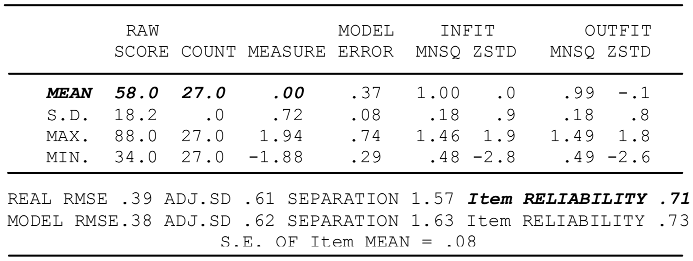
Malaysia has a vibrant ICT and Services industry that is world-class, confirmed by the AT Kearney 2004 and 2005 Offshore Location Attractiveness Index, which ranks Malaysia as the world's third most attractive Shared Services & Outsourcing (SSO) location [16]. Rasch help evaluate small sample size that give 95% confidence level.

A. Summary Statistic

Result shows in Figure 3 with 84% (N = 27) of respondents is a valid response after clean the misfit data. Besides, 77% of an item is measured after clean the invalid item. A total of 2295 data points arising from 27 respondents on 85 items was analyzed. It yields a Chi-Square value of 3493.40. Cronbach-α value is 0.90, which contribute high reliability of raw score for the instrument in measuring the undergraduate problem.

The optimal categorization [15] in which provides the best construct definition, best separates respondents along the variable, and produces the best fit of data to model. Targeting is at 0.81 *logit* (S.E. = 0.13) which referring to *MeanPerson* – *MeanItem* (0.81 *logit* – 0.00 *logit*). Targeting is less than 1 *logit.* Based on a rating scale instrument quality criteria; if targeting < 1 error, then it is good targeting. So, the instrument is on target.

In a mean time, it showed a “Good” reliability (Fisher, 2007) for both item and person reliability. The item is sufficient at 0.71 that above from 0.7 (reliability > 0.7) give a meaning that the instrument has acceptable number of items to measure what is supposedly to be measured in the underpinning theory.

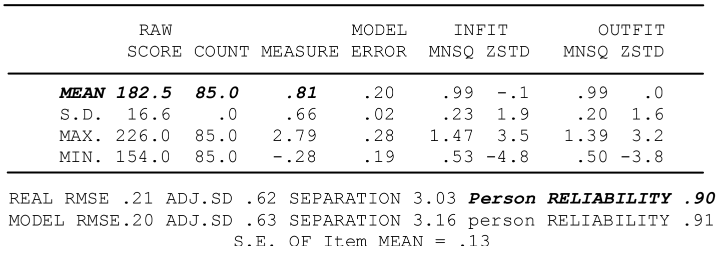


```
                 RAW                          MODEL      INFIT        OUTFIT
                SCORE  COUNT  MEASURE  ERROR   MNSQ  ZSTD    MNSQ  ZSTD

        MEAN     58.0   27.0      .00     .37   1.00    .0     .99   -.1
        S.D.     18.2     .0      .72     .08    .18    .9     .18    .8
        MAX.     88.0   27.0     1.94     .74   1.46   1.9    1.49   1.8
        MIN.     34.0   27.0    -1.88     .29    .48  -2.8     .49  -2.6

REAL RMSE .39 ADJ.SD .61 SEPARATION 1.57 Item RELIABILITY .71
MODEL RMSE.38 ADJ.SD .62 SEPARATION 1.63 Item RELIABILITY .73
                   S.E. OF Item MEAN = .08
```

Fig. 3. Summary of Measured 85 Items

The person reliability is sufficient at 0.90 indicates a high reliability. This gives an indication that the instrument can differentiate the person ability with the difficult past practices in undergraduate study. Furthermore, the item was sufficient to separate the person; PSI 3.08 *logit* into four groups, which match the expectation (CGPA categories). Expected person based on CGPA categories is 0.00-1.99 (D), 2.00 - 2.99 (C), 3.00 – 3.49 (B) and 3.50 – 4.00 (A). No person is in D categories.

MeanItem from Figure 4 is set to an arbitrary 0.00. The instrument where ‘zero-setting’ all items is at 50:50 situation. Error 0.37 (> 0.25) is slightly high. *MeanPerson* give a value of +0.81 logit; Person meet expectation. The person-item map in Figure 6, reveals that there is one poor person is located below *MeanItem* and has low ability with -0.28 *logit*. The Excellent person is at 2.79 *logit*, which above the highest located practice at 1.94 difficulty *logit*. Inspite of the good reliability, more difficult items. However, the items need to be introduced for that gap of 0.95 *logit*

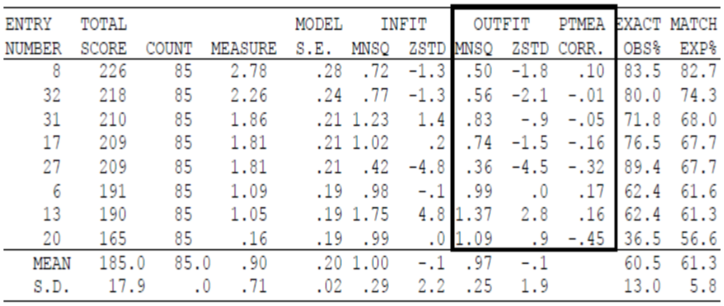


```
                 RAW                          MODEL      INFIT        OUTFIT
                SCORE  COUNT  MEASURE  ERROR   MNSQ  ZSTD    MNSQ  ZSTD

        MEAN    182.5   85.0      .81     .20    .99   -.1     .99    .0
        S.D.     16.6     .0      .66     .02    .23   1.9     .20   1.6
        MAX.    226.0   85.0     2.79     .28   1.47   3.5    1.39   3.2
        MIN.    154.0   85.0     -.28     .19    .53  -4.8     .50  -3.8

REAL RMSE .21 ADJ.SD .62 SEPARATION 3.03 Person RELIABILITY .90
MODEL RMSE.20 ADJ.SD .63 SEPARATION 3.16 person RELIABILITY .91
                   S.E. OF Item MEAN = .13
```

Fig. 4. Summary of Measured 27 Persons

B. Person and Item Fit

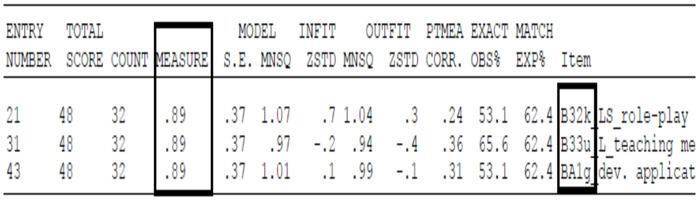
To fit the item into the model, we should first identify the sum of Mean and Standard Deviation (SD) to clean the data based on Point Mean Correlation (PTMEA), Mean Square (MNSQ) and z-standard (ZSTD). If the data is indicated high z-std that bigger than 1.29 logit, it is person misfit. Figure 5 shows person 20, 27, 31, 32 and 13 are misfit with MNSQ > 1.25 *logit* and z-std > ±2 *logit*. Item whose MNSQ is nearer to 1 and z-std nearer to 0 is deemed a better fit. However, Point of Mean (PTMEA) Correlation allow the negative response because the study is to identify the competence graduates students that get low CGPA but has a high skill. Ignore if in between the range of 0.5 < MNSQ < 1.5 and ZSTD ±2 logit.



```
ENTRY   TOTAL                    MODEL     INFIT  |   OUTFIT    PTMEA |EXACT MATCH
NUMBER  SCORE  COUNT   MEASURE   S.E.  MNSQ  ZSTD |MNSQ  ZSTD  CORR.  | OBS%  EXP%
     8    226     85      2.78    .28   .72  -1.3 | .50  -1.8    .10  | 83.5  82.7
    32    218     85      2.26    .24   .77  -1.3 | .56  -2.1   -.01  | 80.0  74.3
    31    210     85      1.86    .21  1.23   1.4 | .83   -.9   -.05  | 71.8  68.0
    17    209     85      1.81    .21  1.02    .2 | .74  -1.5   -.16  | 76.5  67.7
    27    209     85      1.81    .21   .42  -4.8 | .36  -4.5   -.32  | 89.4  67.7
     6    191     85      1.09    .19   .98   -.1 | .99    .0    .17  | 62.4  61.6
    13    190     85      1.05    .19  1.75   4.8 |1.37   2.8    .16  | 62.4  61.3
    20    165     85       .16    .19   .99    .0 |1.09    .9   -.45  | 36.5  56.6
   MEAN   185.0   85.0     .90    .20  1.00   -.1   .97   -.1           60.5  61.3
   S.D.    17.9     .0     .71    .02   .29   2.2   .25   1.9           13.0   5.8
```

Fig. 5. Person Fit

It goes the same to item fit, from 110 item constructed, 25 item deleted. Again, the suggestion is based on Point Mean Correlation (PTMEA), Mean Square (MNSQ) and z-standard (ZSTD). Item whose MNSQ is nearer to 1 and z-std nearer to 0 is deemed a better fit. Remain the same if the item is in between the range of 0.5 < MNSQ < 1.5 and ZSTD ±2 logit. Remain the same to this item whose MNSQ is a measure the same but different group of the item to be measured because of content validity is preserved in Figure 6.

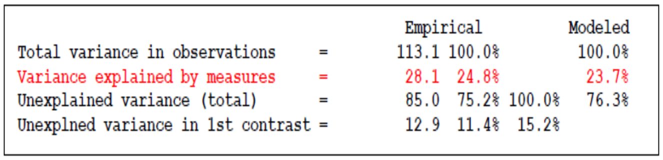


```
ENTRY   TOTAL                 MODEL   INFIT    OUTFIT  PTMEA EXACT MATCH
NUMBER  SCORE COUNT MEASURE  S.E. MNSQ  ZSTD MNSQ  ZSTD CORR. OBS%  EXP%  Item

21       48     32    .89     .37  1.07    .7 1.04    .3   .24  53.1  62.4 B32k LS_role-play
31       48     32    .89     .37   .97   -.2  .94   -.4   .36  65.6  62.4 B33u L_teaching me
43       48     32    .89     .37  1.01    .1  .99   -.1   .31  53.1  62.4 BA1g dev. applicat
```

Fig. 6. Consolidated Item Misfit

C. Unidimensionality

To ensure the measurement is measuring the specific objective, thus, unidimensionality is crucial. Rasch Analysis applies the Principal Component Analysis (PCA) of the residuals to know on how much variance of the instrument measuring that supposedly to measure. The raw variance explained by measures is 24.8% closely match the expected 23.7%. However, the analysis shows that only 20% of unidimensionality requirement minimum. Rasch cut-low point of 40% is not achieved [5, 19]. Nevertheless, the unexplained variance in the 1st contrast of good 11.4% is obtained as tabulated in Figure 7.



```
                                           Empirical           Modeled
Total variance in observations     =       113.1 100.0%         100.0%
Variance explained by measures     =        28.1  24.8%          23.7%
Unexplained variance (total)       =        85.0  75.2% 100.0%   76.3%
Unexplned variance in 1st contrast =        12.9  11.4%  15.2%
```

Fig. 7. Standardized Residual variance (in Eigenvalue units)

VII. PERSON-ITEM MAP

A. Item Analysis

Figure 5 depict the location of the undergraduate practice based on industry personnel experience during their undergraduate study in Requirement Engineering. The location of practice is according to industry personnel ability and difficulty *logit*. Twenty-nine out of eighty-five items discussed in this paper is referring to identify problem in undergraduate study based on industry personnel experience.

Industry personnel experienced on difficulty which involved seventeen (0.59%) items. They are hardly endorsing this item that above Item*Mean*. Eight items in between Item*Max* and Person*Mean* which can be considered as the most difficult item and seventeen industry personnel not experience those items. Final Examination and Motivation from Lecturer measured inline Mean*Item* 0.00 *logit*. This two items show normality of undergraduate Requirement Engineering study. All industry personnel agreed, to easily endorse ten items that below ItemMean.

B. Person Analysis

There are six female and twenty-one male industry personnel randomly picked from Sharing Services Outsourcing was evaluated. The person-item map shown there is five A's person, eighteen B's person and two C's person. We can generalize that Sharing Services Outsourcing hire more industry personnel from B's person rather than A's and C's person.

The most excellent industry personnel is at highest ability 2.79 logit and seen out of the target. The off target person is the one without corresponding items. The poor industry personnel are at -0.28 logit. The difference between PersonMax and PersonMin is 2.51 logit. The difference is slightly over Standard Deviation (SD) of 0.60 logit. This shows the There are 0.56% (N = 15) industry personnel located above MeanPerson and 0.04% (N = 1) is below than minimum MeanPerson. In between A's person, located a C's industry personnel above from MeanPerson.

The highest industry personnel at 2.79 logit is female, 23 - 27 years old, hold a first degree from Software Engineering. Earn CGPA 3.50-3.67 and get requirement grade is in between A- to A. Working as software developer with experience five years and less in software industry. Motivates in successfully finishing project are self-esteem, token from company and responsibility. She involved in 6 to 10 projects in HLE and software industry that involved web-based and multimedia project. She had an experience as developer and system designer.

Other A's person that close to Item*Max* is male. His age is 33 and above. Hold Master in Information Technology (IT). He earn CGPA 3.50-3.67 during his first degree with grade A- to A in Requirement Engineering subject. He had an experience 10 years and above which involved web-based, networking, stand-alone system, others project software development. Experience in doing a software development project is 10 and above project at HLE and software industry. He had an experience as developer, system engineer, system developer, documenter and project manager.

There is B's person below PersonMin is male, age in between 33 and above. He holds First Degree in Computer science. He manages to get CGPA 2.50-2.99 with grade D- to D in Requirement Engineering. He has five and less experience in networking, He had 5 and less experience in HLE and one to five project at industry as a developer.

C's person that range in between A's person is male, age 33 and above, diploma, network, 2.00-2.49,C- to C, system engineer, experience 5 years and less, manufacturing, motivates in successfully finishing project is token from company, rate current learning is average, involved web-based,5 and less project in HLE, 1-5 in working environment, need training after HLE, need tool to capture requirement, average in using internet to finish the project.

Contradict with C's person above PersonMin and B's person is male, age 33 and above, diploma, computer science, 2.00-2.49, C- to C+, software engineer, 5-10 year experience, multimedia project, 5-less project in HLE, 5-10 project experience, experience as developer only.

However, all persons (A's, B's and C's person) agree that they need training after Higher Learning Education (HLE). In increasing the skill among undergraduate students, tool is much recommended to capture requirement from stakeholder. Internet is very useful to finish the project.

VIII. CONCLUSION

As a conclusion, Rasch help to identify missing data. Person and item misfit is managed with careful manner. Content, construct and predictive validity are maintained. Based on Person-Item Map in Figure 8, the unidimensionality is achieved. Research objective to identify the problem during undergraduate study based on industry personnel is achieved and successfully discussed using Rasch Measurement Model. The industry field required more than CGPA achievement. They are facing the real development project which need a skill worker to handle it. The average CGPA needed by the industry is in range 2.50 and above. The skill worker that have the experience in handling the project more than five development project were very highly recommended to field the position in the software industry.

Future work will be discussed on further analysis based on expert recommendation using QSR NVIVO tool to enhance employability skill for Software Engineering undergraduate in HLE.

REFERENCES

[1] Al-Qardhawi, Y. (2003). *Sunnah: Sumber ilmu dan peradaban. Penterjemah*. (M. Firdaus, Ed.) Selangor: International Islamic Thought Malaysia.

[2] Azlinah Mohamed, Azrilah Aziz, Sohaimi Zakaria and Mohd saidfudin Masodi (2008). Appraisal of Course learning Outcomes using Rasch Measurement: A Case Study in Information Technology Education. Conference Proceeding 7th WSEAS International Conference on Artificial Intelligent, Knowledge Engineering and Databases (AIKED ‘08), 20-22 February, 2008, University Cambridge, Cambridge, UK.

[3] Baghaei P. *The Rasch Model as a Construct Validation Tool*,Rasch Measurement Transactions, 2008, 22:1 p. 1145-6 Chapman, A. (2006). Bloom's Taxonomy - Learning Domains. vol. 2007: Businessballs.com.

[4] Bloom B. S. (1956). Taxonomy of Educational Objectives, Handbook I: The Cognitive Domain. New York: David McKay Co Inc.

[5] Conrad et al. (2011). Validation of the Crime and Violence Scale to the Rasch Measurement Model GAIN Methods Report 1.2. p.10.

[6] Fisher, W. P. J. (2007). Rating scale instrument quality criteria. Rasch Measurement Transactions, 21(1), pg 1095.

[7] Gothwal VK., Wright T.A., (2009). Lamoureux E.L., Pesudovs K, *Using Rasch analysis to revisit the validity of the Cataract TyPE Spec instrument for measuring cataract surgery outcomes*, Journal of Cataract Refractive Surgery - Vol 35: p.1509-1517.

[8] Hamzah A. G. & Saidfudin, M. (2009). Modem measurement paradigm in Engineering Education: Easier to read and better analysis using Rasch-based approach. 2009 International Conference on Engineering Education (lCEED 2009), December 7-8, 2009, Kuala Lumpur, Malaysia.

[9] Jamaludin, N. A. A., Sahibuddin, S. & Hidayat, N. H. (2011). Challenges of Project-Based Learning Towards Requirement Engineering. The 10th WSEAS International Conference on Software Engineering, Parallel and Distributed Systems (SEPADS '11)”. University Cambridge, Cambridge, UK.

[10] Jamaludin, N. A. A., Sahibuddin, S. (2011). Development Strategy using Cognitive Domain in e-Requirement Engineering Learning System. IJCSI International Journal of Computer Science Issues, Vol. 8, Issue 5, September 2011. ISSN (Online): 1694-0814. www.IJCSI.org.

[11] Jamaludin, N. A. A., Sahibuddin, S., Jusoff, K., Hidayat, N. H. (2010). Development of a Project-Based Learning Approach in Requirement Engineering. International Journal of Computer Science and Information Security, IJCSIS, Vol. 8, No. 9, pp. 6.

[12] Leedy, P., Ormond, J. (2011). *Practical Research:Planning and Design*, Pearson.

[13] Linacre J.M. (2001). *Category, Step and Threshold: Definitions & Disordering*, Rasch Measurement Transactions, 15:1 p.794.

[14] Linacre, J. M. (2011). Winsteps Rasch Measurement Version 3.71. Winstep.com.

[15] Lopez W (1996) Communication Validity and Rating Scales. Rasch Measurement Transactions 10:1 p. 482-3)

[16] Multimedia Development Corporation (MDeC). http://www.mscmalaysia.my, retrieved 23rd November 2011).

[17] Pohl, M. (2000). Learning to Think, Thinking to Learn: Models and Strategies to Develop a Classroom Culture of Thinking. Cheltenham, Vic.: Hawker Brownlow.

[18] Ramli M. I. (2001). Disclosure in annual reports: an agency theoretic perspective in an International setting. Unpublished Doctoral Dissertation. University of Plymouth, United Kingdom.

[19] Reckase, M.D. (1979). Unifactor Latent Trait Models Applied to Multifactor Tests: Results and Implications, Journal of Educational Statistics, Vol. 4, No. 3, pp. 207-230.

[20] Saidfudin, M., Azlinah M., Azrilah A. A. , NorHabibah, A., Hamzah A. G. & Sohaimi, Z., (2008). Application of Rasch Model in validating the construct of measurement instrument", in International Journal of Education and Information Technologies, Issue 2, Volume 2,. pp. 105-112.

[21] Saidfudin, M., Hamza, A. G., Razimah, A. & Rozeha, A. (2008). Application of Rasch-based ESPEGS Model in Measuring Generic Skills of Engineering Students: A New Paradigm, WSEAS Transactions on Advances in Engineering Education, Issue 8 Vo1.5, WSEAS Press. pp. 591-602, August 2008.

[22] Solemon, B., Sahibuddin, S., Ghani, A.A.A. (2009) Requirements engineering problems and practices in software companies: An industrial survey. *International Conference on Advanced Software Engineering and Its Applications, ASEA 2009 Held As Part of the Future Generation Information Technology Conference*, pp. 70-77. FGIT 2009, Jeju Island, Korea, December 10-12.

[23] Wright B. D. and Mok, M. M. C. (2004). An overview of the family of rasch measurement models," in Introduction to Rasch Measurement: Theory, Models, and Applications, J. Everett V.Smith and R. M.Smith, Eds. p. 979.

[24] http://www.uky.edu/~kdbrad2/AERA2005SupplyDemandPaper.pdf;

[25] http://www.wseas.us/e-library/conferences/2010/Penang/MMF/MMF-25.pdf;

[26] http://www.docstoc.com/docs/55062758/Rasch-Workshop-Booklet---Structu;

AUTHORS PROFILE

Nor Azliana Akmal Jamaludin is a Lecturer for Bachelor of Science Software Engineering and Head of Developer for Master Degree Software Engineering program at Universiti Industri Selangor. She received the Master Degree in Computer Science (Real-Time Software Engineering) from Advanced Informatics School (formerly known as Centre for Advanced Software Engineering (CASE), Universiti Teknologi Malaysia, in 2004. Currently, she is doing her Doctorate in Computer Science, specialize in Software Engineering at UTM and supervise by Prof. Dr. Shamsul Sahibuddin. She is a member of Malaysian Software Engineering Interest Group, Malaysia. Her field of expertise is in software requirement, analysis, system integration, and e-learning. Her current research interest is on techniques that can enhance skill among Software Engineering undergraduate of higher institutions in Malaysia using machine learning system.

Shamsul Sahibuddin received the PhD in Computer Science at Aston University, Birmingham, United Kingdom in 1998 and Master Science (Computer Science) at Central Michigan University, Mt. Pleasant, Michigan in 1988. He is currently a Dean / Professor at Advanced Informatics School, Universiti Teknologi Malaysia (UTM) and has been a Member of the Program Committee for Asia-Pacific Conference on Software Engineering. His field of expertise is in software requirement, software modelling technique, software development and software process improvement. He has been very active in scholarly journals-book writing and publishing.

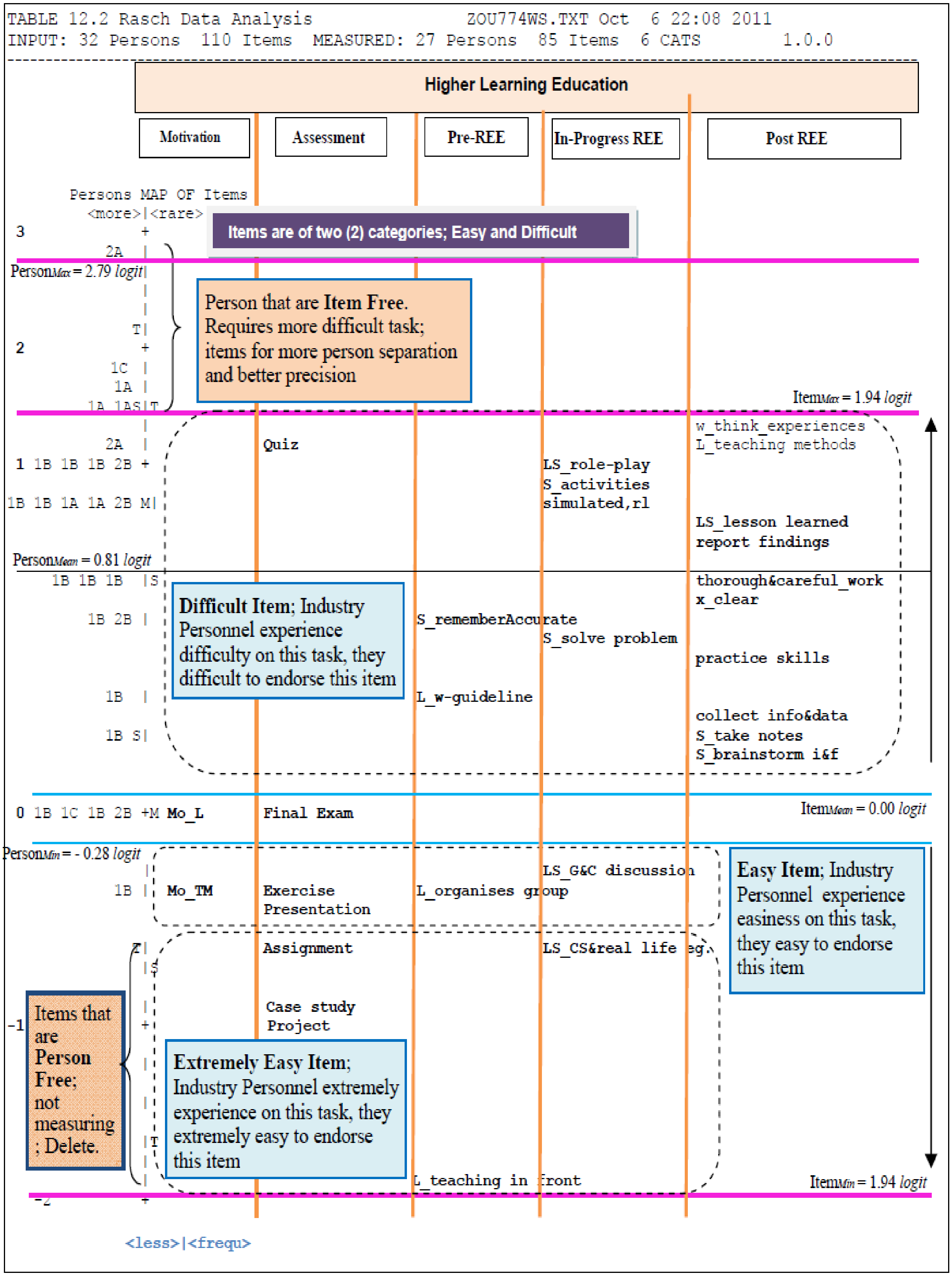


Fig. 8. Person-Item Map for Requirement Engineering Education (REE)

A Novel Steganography Method for Hiding BW Images into Gray Bitmap Images via k-Modulus Method

Firas A. Jassim
Management Information Systems Department
Faculty of Administrative Sciences
Irbid National University
Irbid 2600, Jordan
Email: firasajil@yahoo.com

***Abstract*—This paper is to create a pragmatic steganographic implementation to hide black and white image which is known as stego image inside another gray bitmap image that known as cover image. First of all, the proposed technique uses k-Modulus Method (K-MM) to convert all pixels within the cover image into multiples of positive integer named k. Since the black and white images can be represented using binary representation, i.e. 0 or 1. Then, in this article, the suitable value for the positive integer k is two. Therefore, each pixel inside the cover image is divisible by two and this produces a reminder which is either 0 or 1. Subsequently, the black and white representation of the stego image could be hidden inside the cover image. The ocular differences between the cover image before and after adding the stego image are insignificant. The experimental results show that the PSNR values for the cover image are very high with very small Mean Square Error.**

Keywords*—*Image steganography, Information hiding, Security, k-Modulus Method.

I. Introduction

Recently, there has been great interest in image steganography and its implementation in the filed of information security. The word steganography has been originated from the Greek words stegos which means covered and graphia which means writing. Actually, the word steganography refers to the process of concealing secret data into any kind of media such as image, video, or audio [12]. The essential concept in every data hiding method is the feebleness of human perception such as vision, listening, and hearing. It must be differentiated between steganography and cryptography because both of them are used to hide secret data. The basic dissimilarity between steganography and cryptography is that cryptography focuses on preserving the contents of the message confidential. In the other hand, steganography concentrates on preserving the existence of a message confidential at the first place [13]. Hence, if the notion is to conceal the existence of the secret message, then the method of steganography is preferred [13]. Also, it must be differentiated between steganography and watermarking because of the common confusion between them. The essential difference between steganography and watermarking is the absence of an adversary. In watermarking there is an active adversary that would try to forge the watermarks. On the contrary, in steganography there is no such an active adversary [4]. Recently, there are several methods have been proposed for image based steganography by many authors. The simplest and the most common method is the Least Significant Bit (LSB) which replaces the least significant bits from the right of a pixel to hide the information [6]. Furthermore, there are wide manifold techniques with their own pros and cons were developed in steganography. Many authors have researched the method of hiding information inside image via steganographic technique [3][16][15]. Also, an excellent theoretical background about steganography could be found in [2].

The organization of this article is as follows: In section II, a brief discussion about k-Modulus Method was presented. The proposed steganographic technique was discussed in Section III. In addition, experimental results and conclusions are presented in Sections IV and V, respectively.

II. k-Modulus Method

The first origination of k-Modulus Method was as an image compression method that was proposed by [7]. The basic concept of Five Modulus Method is to transform the whole pixels inside the original image into multiples of five. After that, a generalization of Five Modulus Method was established and named k-Modulus Method (k-MM) where k is any positive integer [6]. The k-Modulus method as an image transformation method proven its ability to mimics the original image before the transformation. In other words, the human eye can not differentiate between the original image and the transformed k-Modulus Method image, fig 1.

Actually, k-Modulus Method transform could be considered as a suitable carrier to convey data inside. The conveyed data could be hidden inside the non-divisible integers of k. Therefore, any pixel that is not divisible by k with reminder zero, this implies that there is hidden information in this pixel and so on. Since the aim of this research is to hide black and white images inside gray image, then the demand for two binary representations is arise. Therefore, choosing k is equal to two is compulsory for two reasons. The first reason is to hide binary data as 0 or 1. The second reason is that whenever there is an increase in k, the transformed image could be distorted. Hence, the best expedient that could be used to preserve the original pixel values as could as possible.

As the topic of k-Modulus method considered as a new born topic, it has many applications is image processing field. According to [9], the embedding of k-Modulus Method into JPG image format to increase the compression ratio was discussed. Moreover, in accordance to [10], increasing the compression ratio in PNG file format was also obtained by the assistant of k-Modulus Method. Finally, the implementation of k-Modulus Method in steganography in two approaches. The first approach, is by hiding text in an image using Five Modulus Method which was discussed by [11]. The second approach which was researched by [8], was the first seed to hide small size image into another larger image.

(a) Original Lena (b) Transformed Lena

Fig. 1: k-Modulus Method Transformation (k=2) with PSNR=50.7787

III. Proposed Steganography Technique

As mentioned previously, k-Modulus method could be treated as a robust host to convey information. The proposed technique starts with transforming the original image into gray scale image. This step is to reduce image size because the color bitmap image is three times higher in size than gray scale image with the same dimensions. Hence, in this research, a gray scale bitmap image could be used as a cover image. The resulted gray scale bitmap image is now ready for the k-Modulus Method transformation with k=2, as mentioned in the previous section. Now, the numerical values for all the pixels within the gray scale bitmap image are of multiples of two. Mathematically speaking, the reminder of dividing any positive integer number by two yields either zero or one. Fortunately, this may be treated as a robust host to accommodate images with binary representation, i,.e. black and white images. Concerning stego (secret) image, it must be a binary (black and white) image. Obviously, the black and white images are the images that consist of zeros and ones. Till this step, two images were constructed which are the cover gray scale bitmap image and the stego black and white image. The final step consists of adding the cover and the stego images together. The resulted image pixels are either multiples of two or multiples of two plus one. If the reminder of the pixel is zero, then it is treated as zero representation, otherwise it is one. hence, one can easily extract the hidden secret black and white image. The reason for choosing black and white image in the proposed steganographic technique is to hide images that contain an essential information such as maps and geometric line. Now, an illustrative example will be discussed to clarify the proposed technique.

Firstly, an arbitrary 10×10 block from Lena gray scale bitmap image will be considered as a cover image, table I. Secondly, the k-Modulus method transformation will be applied to the 10×10 block, tableII. Thirdly, a random 10×10 block from any black and white image will be used as a stego image, table III. Finally, the addition of the cover and stego images could be presented in table IV. Clearly, the resulted block could be divisible by two and produces two reminders which are zero and one. Obviously, the resulted reminders could be formulate the stego (secret) image exactly.

TABLE I: 10×10 Block from Original Lena image

93	95	100	96	98	100	101	97	99	102
96	95	92	100	94	97	97	93	102	106
97	99	98	101	98	98	95	96	99	100
96	95	95	96	100	97	98	97	100	103
100	95	98	97	100	104	98	97	101	101
92	97	93	100	99	98	99	96	101	102
98	96	95	94	99	101	97	100	102	103
95	94	96	98	101	102	95	101	105	100
101	98	99	100	92	104	101	102	105	105
100	100	105	103	102	99	98	100	103	104

TABLE II: 10×10 Block from Transformed Lena image

92	94	100	96	98	100	100	96	98	102
96	94	92	100	94	96	96	92	102	106
96	98	98	100	98	98	94	96	98	100
96	94	94	96	100	96	98	96	100	102
100	94	98	96	100	104	98	96	100	100
92	96	92	100	98	98	98	96	100	102
98	96	94	94	98	100	96	100	102	102
94	94	96	98	100	102	94	100	104	100
100	98	98	100	92	104	100	102	104	104
100	100	104	102	102	98	98	100	102	104

TABLE III: 10×10 Block from black and white Gadeer (Stego) image

1	1	1	0	0	0	0	0	0	0
1	1	1	0	0	0	0	0	0	0
1	1	1	1	0	0	0	0	0	0
1	1	1	1	0	0	0	0	0	0
1	1	1	1	0	0	0	0	0	0
1	1	1	1	0	0	0	0	0	0
1	1	1	0	0	0	0	0	0	0
1	1	1	0	0	0	0	0	0	0
1	1	0	0	0	0	0	0	0	0
1	0	0	0	0	0	0	0	0	0

TABLE IV: 10×10 Block obtained by adding Gadeer to Lena

93	95	101	96	98	100	100	96	98	102
97	95	93	100	94	96	96	92	102	106
97	99	99	101	98	98	94	96	98	100
97	95	95	97	100	96	98	96	100	102
101	95	99	97	100	104	98	96	100	100
93	97	93	101	98	98	98	96	100	102
99	97	95	94	98	100	96	100	102	102
95	95	97	98	100	102	94	100	104	100
101	99	98	100	92	104	100	102	104	104
101	100	104	102	102	98	98	100	102	104

On the sender side, the sender will send the constructed cover image that was presented in table IV. On the receiver side, the recipient will divide the whole image by two and store the reminder in a new array. The resulted array is exactly the stego black and white image. It must be mentioned that, the reason for using bitmap is that, bitmap image does not modify its structure since it is a lossless compression technique. In other words, there was a need to preserve the constructed image values exactly. Therefore, a lossless method of compression was proposed. On the contrary, if lossy compression was used instead of lossless then a completely different results will be obtained. In fact, lossy compression does not be appropriate for the proposed technique because it is irreversible. In other words, one can not retrieve the original pixels when using lossy compression. Hence, JPG compression is not suitable for the proposed steganographic techniqe.

IV. Experimental Results

In this section, both a practical and visual evidences have been implemented to support the proposed technique. Actually, the proposed technique in this article has been implemented to four 512×512 test images with different spatial and frequency characteristics as cover images, fig. 2. Since the proposed technique hide stego image with the same size of the cover image, then six 512×512 black and white stego (secret) images, fig. 3. According to [5], two error measures have been utilized to evaluate the differences between the original and the cover images. The first error measure is the widely and simplest one which is the mean square error (MSE) that can be expressed as:

$$MSE = \frac{1}{NM}\sum_{x=1}^{N}\sum_{y=1}^{M}[f(x,y) - g(x,y)]^2 \qquad (1)$$

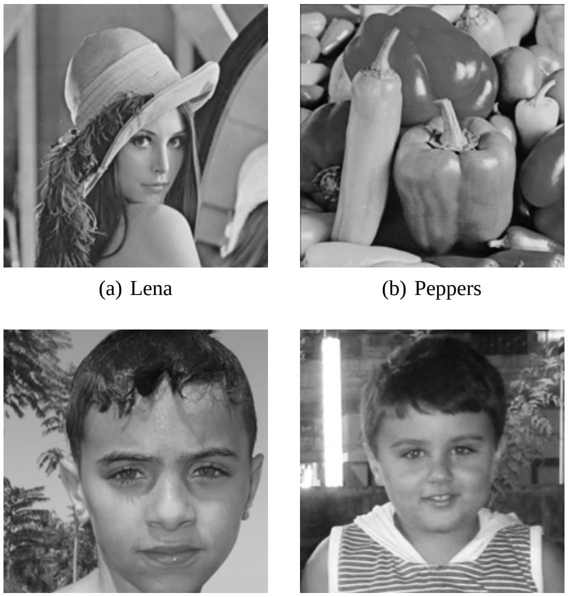
The second error metric is the peak signal to noise ratio (PSNR) which is the most preferable measure that computes the dissimilarities between images in most of the image processing fields [5]. The mathematical formula for the PSNR is as follows:

$$PSNR = 20 \cdot log_{10}\frac{MAX}{\sqrt{MSE}} \qquad (2)$$

The computed error metrics for the proposed technique have been evaluated and introduced in tables V and VI, respectively. Clearly, the mean square error values in table V seem to be very small which implies the closer the values between the original and the constructed cover images. This may consolidate the claim about the proposed technique in which that there are almost tinny dissimilarities that can not be distinguished by the human eye. Hence, any adversary can not notice or even feel that the cover image contain any secret information. Additionally, the PSNR values have been calculated and presented in table VI. The higher the value of PSNR, the more quality the stego image will have. In accordance to [1] [14], the reasonable range for the acceptable PSNR values are between 30 and 50. Therefore, the computed PSNR values are highly tolerable because all the computed PSNR values are higher than 50. Consequently, this support the proposed technique to hide black and white image confidentially in any gray bitmap images.

V. Conclusion

In this paper, a novel confidential steganographic technique was proposed. The essential conclusion that comes from the proposed technique is the high confidentiality and conceal ability to the embedded information inside the cover image. However, the proposed technique is very suitable in sending black and white images like diagrams, line chart, secret maps, etc. Moreover, the file size for the original image will remain constant even after embedding the secret stego image inside. This is one of the important benefits of the proposed technique that helps not to suspect about the cover image and its contents by any adversary. practically, experiment results have demonstrated that the proposed technique produces high PSNR values that makes no doubt that the suggested steganographic technique does not affect the cover image at all.



(a) Lena

(b) Peppers

(c) Saif

(d) Aws

Fig. 2: Four cover Images

TABLE V: Mean Square Error

	Lena	Peppers	Saif	Aws
Gadeer	0.4977	0.4999	0.4977	0.4980
Noon	0.4993	0.4996	0.4992	0.4918
Text1	0.4996	0.5004	0.4992	0.4945
Text2	0.4988	0.4990	0.4994	0.4924
Mask5	0.4981	0.5000	0.4984	0.4945
Barcode	0.5001	0.5006	0.5003	0.5016

TABLE VI: PSNR values for the cover images

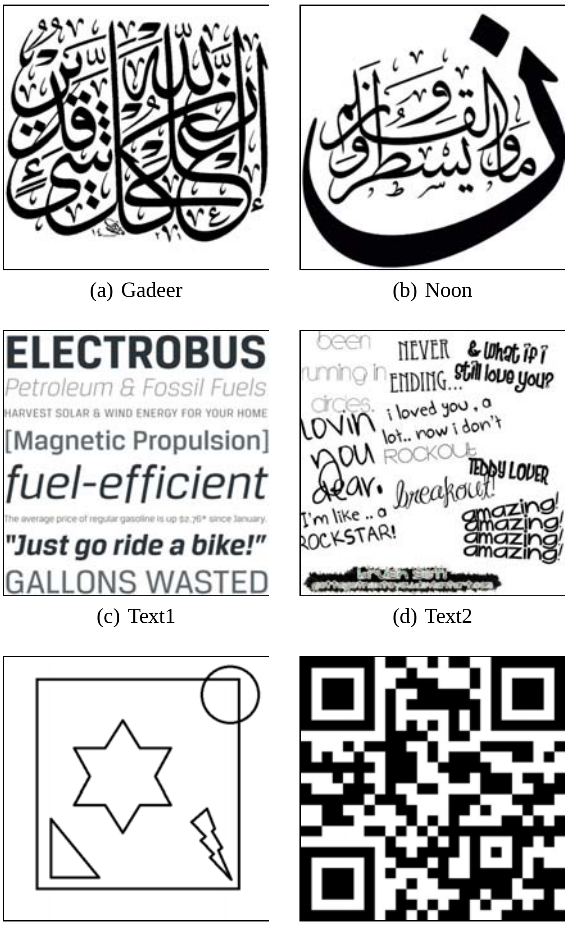
	Lena	Peppers	Saif	Aws
Gadeer	50.8132	50.4329	50.9543	51.1589
Noon	50.7993	50.4355	50.9416	51.2131
Text1	50.7967	50.4282	50.9411	51.1894
Text2	50.8038	50.4400	50.9392	51.2075
Mask5	50.8099	50.4315	50.9484	51.1891
Barcode	50.7926	50.4268	50.9318	51.1270

References

[1] M. Barni, *Document and Image Compression*. Taylor & Francis, 2006.

[2] C. Cachin, “An information-theoretic model for steganography,” *Inf. Comput.*, vol. 192, no. 1, pp. 41–56, Jul. 2004.

[3] C. C. Chang, J. Y. Hsiao, and C. C. S., “Finding optimal least significant bit substitution in image hiding by dynamic programming strategy,” *Pattern Recognition*, vol. 36, pp. 1583–1595, 2003.

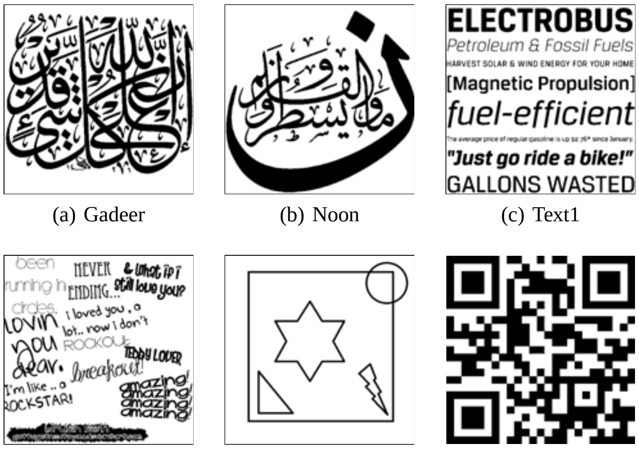


(a) Gadeer (b) Noon

(c) Text1 (d) Text2

(e) Shapes (f) Barcode

Fig. 3: Six stego Images



(a) Gadeer (b) Noon (c) Text1

(d) Text2 (e) Shapes (f) Barcode

Fig. 4: Extracted stego images from Lena as a cover image

[4] A. Cheddad, J. Condell, K. Curran, and P. Mc Kevitt, “Review: Digital image steganography: Survey and analysis of current methods,” *Signal Process.*, vol. 90, no. 3, pp. 727–752, Mar. 2010.

[5] R. C. Gonzalez and R. E. Woods, *Digital image processing*, 3rd ed. Prentice Hall, 2008.

[6] F. A. Jassim, “k-modulus method for image transformation,” *International Journal of Advanced Computer Science and Applications*, vol. 4, no. 2, pp. 267–271, 2013.

[7] F. A. Jassim, “Five modulus method for image compression,” *Signals and Image Processing: An International Journal*, vol. 3, no. 5, pp. 19–28, 2012.

[8] F. A. Jassim, “Hiding image in image by five modulus method for image steganography,” *Journal of computing*, vol. 5, no. 2, pp. 21–25, 2013.

[9] F. A. Jassim, “Image compression by embedding five modulus method into jpeg,” *Signals and Image Processing: An International Journal*, vol. 4, no. 2, pp. 31–39, 2013.

[10] F. A. Jassim, “Increasing compression ratio in png image by k-modulus method for image transformation,” *International Journal of Advanced Research in Computer Science and Software Engineering*, vol. 3, no. 6, pp. 45–52, 2013.

[11] F. A. Jassim, “A novel steganography algorithm for hiding text in image using five modulus method,” *International Journal of Computer Applications*, vol. 72, no. 17, pp. 39–44, 2013.

[12] N. Johnson and S. Jajodia, “Exploring steganography: Seeing the unseen,” *IEEE Computer Society*, vol. 31, no. 2, pp. 26–34, 1998.

[13] H. Wang and S. Wang, “Cyber warfare: Steganography vs. steganalysis,” *Communications of the ACM*, vol. 47, no. 10, 2004.

[14] S. Welstead, *Fractal and Wavelet Image Compression Techniques*. SPIE, 1999.

[15] A. Westfeld, “F5-a steganographic algorithm high capacity despite better steganalysis,” *Lecture Notes in Computer Science*, vol. 2137, pp. 289–302, 2001.

[16] P. C. Wu and W. H. Tsai, “A steganographic method for images by pixel-value differencing,” *Pattern Recognition Letters*, vol. 24, pp. 1613–1626, 2003.