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

Volume 6 Issue 2
February 2015

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

www.ijacsa.thesai.org

IJACSA

WHERE WISDOM SHARES

INTERNATIONAL JOURNAL OF
ADVANCED COMPUTER SCIENCE AND APPLICATIONS

THE SCIENCE AND INFORMATION ORGANIZATION

www.thesai.org | info@thesai.org

Editorial Preface

From the Desk of Managing Editor…

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

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

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

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

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

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

Thank you for Sharing Wisdom!

Managing Editor
IJACSA
Volume 6 Issue 2 February 2015
ISSN 2156-5570 (Online)
ISSN 2158-107X (Print)

Editorial Board

Reviewer Board Members

- **Abassi Ryma**
  Higher Institute of Communications Studies of Tunis , Iset'com
- **Abbas Karimi**
  Islamic Azad University Arak Branch
- **Abdelghni Lakehal**
  Université Abdelmalek Essaadi Faculté Polydisciplinaire de Larache Route de Rabat, Km 2 - Larache BP. 745 - Larache 92004. Maroc.
- **Abdel-Hameed A. Badawy**
  Arkansas Tech University
- **Abdur Rashid Khan**
  Gomal Unversity
- **Abeer Mohamed ELkorany**
  Faculty of computers and information, Cairo Univesity
- **ADEMOLA ADESINA**
  University of the Western Cape
- **Aderemi A. Atayero**
  Covenant University
- **Ahmed S.A AL-Jumaily**
  Ahlia University
- **Ahmed Boutejdar**
- **Ahmed Nabih Zaki Rashed**
  Menoufia University
- **Akbar Hossain**
- **Akram Belghith**
  University Of California, San Diego
- **Albert Alexander S**
  Kongu Engineering College
- **Alci-nia Zita Sampaio**
  Technical University of Lisbon
- **Alexandre Bouënard**
  Sensopia
- **Ali Ismail Awad**
  Luleå University of Technology
- **Amitava Biswas**
  Cisco Systems
- **Anand Nayyar**
  KCL Institute of Management and Technology, Jalandhar
- **Andi Wahju Rahardjo Emanuel**
  Maranatha Christian University
- **Andrews Samraj**
  Mahendra Engineering College
- **Anirban Sarkar**
  National Institute of Technology, Durgapur
- **Antonio Formisano**
- **Anuranjan misra**
  Bhagwant Institute of Technology, Ghaziabad, India
- **Appasami Govindasamy**
- **Arash Habibi Lashkari**
  University Technology Malaysia(UTM)
- **Aree Ali Mohammed**
  Directorate of IT/ University of Sulaimani
- **Aris Skander Skander**
  Constantine 1 University
- **Ashok Matani**
  Government College of Engg, Amravati
- **Ashraf Mohammed Iqbal**
  Dalhousie University and Capital Health
- **Ashraf Hamdy Owis**
  Cairo University
- **Asoke Nath**
  St. Xaviers College(Autonomous), 30 Park Street, Kolkata-700 016
- **Ayad Ghany Ismaeel**
  Department of Information Systems Engineering-Technical Engineering College-Erbil Polytechnic University, Erbil-Kurdistan Region- IRAQ
- **Ayman EL-SAYED**
  Computer Science and Eng. Dept., Faculty of Electronic Engineering, Menofia University
- **Babatunde Opeoluwa Akinkunmi**
  University of Ibadan
- **Badre Bossoufi**
  University of Liege
- **BASANT KUMAR VERMA**
  JNTU
- **Basil Hamed**
  Islamic University of Gaza
- **Basil M Hamed**
  Islamic University of Gaza
- **Bhanu Prasad Pinnamaneni**
  Rajalakshmi Engineering College; Matrix Vision GmbH
- **Bharti Waman Gawali**
  Department of Computer Science & information T

CONTENTS

Paper 1: Effective Strategies for ROI and Image Matching
Authors: Dr. Khaled M. G. Noaman, Dr. Jamil Abdulhamid M. Saif
PAGE 1 – 4

Paper 2: kEFCM: kNN-Based Dynamic Evolving Fuzzy Clustering Method
Authors: Shubair Abdulla, Amer Al-Nassiri
PAGE 5 – 13

Paper 3: Constraint on Repair Resources, Optimal Number of Repairers and Optimal Size of a Serviced System
Authors: Marin Todinov
PAGE 14 – 23

Paper 4: Supporting Self-Organization with Logical-Clustering Towards Autonomic Management of Internet-of-Things
Authors: Hasibur Rahman, Theo Kanter, Rahim Rahmani*
PAGE 24 – 33

Paper 5: Data Center Governance Information Security Compliance Assessment Based on the Cobit Framewok
Authors: Andrey Ferriyan, Jazi Eko Istiyanto
PAGE 34 – 36

Paper 6: Intelligent Traffic Information System Based on Integration of Internet of Things and Agent Technology
Authors: Hasan Omar Al-Sakran
PAGE 37 – 43

Paper 7: Development of a Decision Support System for Handling Health Insurance Deduction
Authors: Shakiba Khademolqorani, Ali Zeinal Hamadani
PAGE 44 – 51

Paper 8: A Multi-Label Classification Approach Based on Correlations Among Labels
Authors: Raed Alazaidah, Fadi Thabtah, Qasem Al-Radaideh
PAGE 52 – 59

Paper 9: Developing Software Bug Prediction Models Using Various Software Metrics as the Bug Indicators
Authors: Varuna Gupta, Dr. N. Ganeshan, Dr. Tarun K. Singhal
PAGE 60 – 65

Paper 10: The Effects of Different Congestion Management Algorithms over Voip Performance
Authors: Szabolcs Szilágyi
PAGE 66 – 70

Paper 11: Study of Gamification Effectiveness in Online e-Learning Systems
Authors: Ilya V. Osipov, Evgeny Nikulchev, Alex A. Volinsky, Anna Y. Prasikova
PAGE 71 – 77

Paper 12: The Real-Time Research of Optimal Power Flow Calculation in Reduce Active Power Loss Aspects of Power Grid
Authors: Yuting Pan, Yuchen Chen, Zhiqiang Yuan, Bo Liu
PAGE 78 – 82

Paper 13: Assessment of Potential Dam Sites in the Kabul River Basin Using GIS
Authors: RASOOLI Ahmadullah, KANG Dongshik
PAGE 83 – 89

Paper 14: The Examination of Using Business Intelligence Systems by Enterprises in Hungary
Authors: Peter Sasvari
PAGE 90 – 96

Paper 15: Sentiment Analysis Based on Expanded Aspect and Polarity-Ambiguous Word Lexicon
Authors: Yanfang Cao, Pu Zhang, Anping Xiong
PAGE 97 – 103

Paper 16: GPS-Based Daily Context Recognition for Lifelog Generation Using Smartphone
Authors: Go Tanaka, Masaya Okada, Hiroshi Mineno
PAGE 104 – 112

Paper 17: Age Estimation Based on AAM and 2D-DCT Features of Facial Images
Authors: Asuman Günay, Vasif V. Nabiyev
PAGE 113 – 119

Paper 18: Personal Health Book Application for Developing Countries
Authors: Seddiq Alabbasi, Andrew Rebeiro-Hargrave, Kunihiko Kaneko, Ashir Ahmed, Akira Fukuda
PAGE 120 – 128

Paper 19: En-Route Vehicular Traffic Optimization
Authors: Saravanan M, Ashwin Kumar M
PAGE 129 – 138

Paper 20: Developement of Bayesian Networks from Unified Modeling Language for Learner Modelling
Authors: ANOUAR TADLAOUI Mouenis, AAMMOU Souhaib, KHALDI Mohamed
PAGE 139 – 144

Paper 21: High Accuracy Arabic Handwritten Characters Recognition Using Error Back Propagation Artificial Neural Networks
Authors: Assist. Prof. Majida Ali Abed, Assist. Prof. Dr. Hamid Ali Abed Alasad
PAGE 145 – 152

Paper 22: The Parents' Perception of Nursing Support in their Neonatal Intensive Care Unit (NICU) Experience
Authors: Amani F. Magliyah, Muhamamd I. Razzak
PAGE 153 – 158

Paper 23: Hybrid PSO-MOBA for Profit Maximization in Cloud Computing
Authors: Dr. Salu George
PAGE 159 – 163

Paper 24: Semantic Web Improved with the Weighted IDF Feature
Authors: Mrs. Jyoti Gautam, Dr. Ela Kumar
PAGE 164 – 173

Paper 25: Consuming Web Services on Android Mobile Platform for Finding Parking Lots
Authors: Isak Shabani, Besmir Sejdiu, Fatushe Jasharaj
PAGE 174 – 180

Paper 26: Improvement of Control System Performance by Modification of Time Delay
Authors: Salem Alkhalaf
PAGE 181 – 185

Paper 27: Use of Non-Topological Node Attribute Values for Probabilistic Determination of Link Formation
Authors: Abhiram Gandhe, Parag Deshpande
PAGE 186 – 191

Paper 28: Different Classification Algorithms Based on Arabic Text Classification: Feature Selection Comparative Study
Authors: Ghazi Raho, Riyad Al-Shalabi, Ghassan Kanaan, Asma'aNassar
PAGE 192 – 195

Paper 29: Implementation of ADS Linked List Via Smart Pointers
Authors: Ivaylo Donchev, Emilia Todorova
PAGE 196 – 203

Paper 30: A Review on Feature Extraction and Feature Selection for Handwritten Character Recognition
Authors: Muhammad 'Arif Mohamad, Haswadi Hassan, Dewi Nasien, Habibollah Haron
PAGE 204 – 212

Paper 31: Resource Provisioning in Single Tier and Multi-Tier Cloud Computing: "State-of-the-Art"
Authors: Marwah Hashim Eawna, Salma Hamdy Mohammed, El-Sayed M. El-Horbaty
PAGE 213 – 217

Paper 32: Improving Web Movie Recommender System Based on Emotions
Authors: Karzan Wakil, Rebwar Bakhtyar, Karwan Ali, Kozhin Alaadin
PAGE 218 – 226

Paper 33: Service Design for Developing Multimodal Human-Computer Interaction for Smart Tvs
Authors: Sheng-Ming Wang, Cheih-Ju Huang
PAGE 227 – 234

Paper 34: A General Model for Similarity Measurement between Objects
Authors: Manh Hung Nguyen, Thi Hoi Nguyen
PAGE 235 – 239

Paper 35: Analysis of Significant Factors for Dengue Infection Prognosis Using the Random Forest Classifier
Authors: A.Shameem Fathima, D.Manimeglai
PAGE 240 – 245

Paper 36: Confinement for Active Objects
Authors: Florian Kammuller
PAGE 246 – 261

Paper 37: Processing the Text of the Holy Quran: a Text Mining Study
Authors: Mohammad Alhawarat, Mohamed Hegazi, Anwer Hilal
PAGE 262 – 267

Paper 38: SOCIA: Linked Open Data of Context behind Local Concerns for Supporting Public Participation
Authors: Shun Shiramatsu, Tadachika Ozono, Toramatsu Shintani
PAGE 268 – 277

Paper 39: Timed-Release Certificateless Encryption
Authors: Toru Oshikiri, Taiichi Saito
PAGE 278 – 284

Paper 40: Vehicle Embedded Data Stream Processing Platform for Android Devices
Authors: Shingo Akiyama, Yukikazu Nakamoto, Akihiro Yamaguchi, Kenya Sato, Hiroaki Takada
PAGE 285 – 294

Paper 41: Ontology-based Change Propagation in Shareable Health Information Applications
Authors: Anny Kartika Sari, Wenny Rahayu
PAGE 295 – 305

Paper 42: Similarity Calculation Method of Chinese Short Text Based on Semantic Feature Space
Authors: Liqiang Pan, Pu Zhang, Anping Xiong
PAGE 306 – 310

Effective Strategies for ROI and Image Matching

Dr. Khaled M. G. Noaman
Associate Professor
Department of Distance Learning
Deanship of E-Learning and Distance Learning,
Jazan University
Jazan, Kingdom of Saudi Arabia (KSA)

Dr. Jamil Abdulhamid M. Saif
Associate Professor
Information Systems Department
Community College of Bisha, Bisha University
Bisha, Kingdom of Saudi Arabia (KSA)

***Abstract*—The paper presents an exceptional four matching strategies: systematic, random, gradient and simulated annealing using diferent metrics. We consider two kinds of image matching algorithms. The first one oriented on the whole image matching where we compare corresponding pixels or chosen image characteristics. The second one is oriented on finding the region in the target image (region of interest ROI) , which match best the ROI given in the template image. For our experiments we take the list of target images, directly from the atlas, and a subset of these images as the template images.**

Keywords—systematic; random; gradient; simulated annealing

I. INTRODUCTION

Presently digital image processing has a broad spectrum of applications, such as multimedia systems, business systems, monitoring and inspection systems, archiving systems. Architectures of such systems are much complex (see Fig. 1). In spite of digitisation, storage, transmission, and display operations, extra functions are considered. They are as follows: image data compression and representation, image enhancement and reconstruction, image indexing retrieval and matching, etc. and they are executed on application oriented servers.

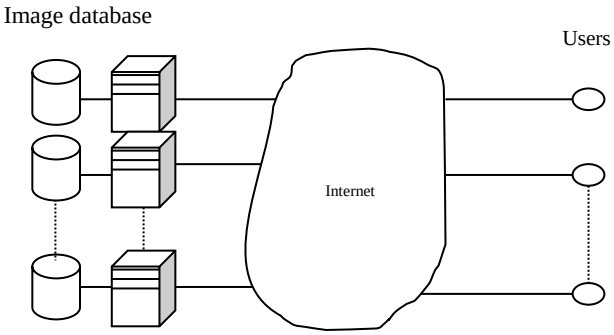


Fig. 1. Modem digital image processing system

In medical applications clinical diagnosis as well as evaluation of therapy is often supported by image processing systems. Such techniques as SPECT (Single Photon Emission Computed Tomography), PET (Positron Emission Tomography), MRS (Magnetic Resonance Spectroscopy), MRI (Magnetic Resonance Imaging), or ultrasound and X-ray scanning are largely used and developed [1,2]. The Internet creates a new possibilities for medicine diagnosis. Application of multiedia systems provides a real-time medical monitoring multi-party consultations and distance collaborations. Examples of such solutions are the following systems:

1) MedNet - used in brain surgery [3],

2) Telematic microscopy system used in diagnostics of histopathology [4],

3) Medinet - used in diagnostics of teleradiology [5].

The rest of this paper is organized as follows: section 2 presents the matching problem and defines the similarity for Whole image and region-based image matching. In section 3, experimental results are presented and discussed, also the effectiveness of our proposed method are discussed. The conclusion and the proposal of future works are given in Section4.

II. IMAGE MATCHING PROBLEM

A digital image (I(m,n), m, n-integers) [6,7,8] is usually the result of discretization process of a continuous image function I(x, y) x, y ∈ R, and it is stored in a computer memory as a two dimensional array A, where A=[A(m, n)], m= 1, 2, ..., M, n=1, 2, ..., N; i.e:

We limit our considerations to the discrete image describing by two dimensional array A. However, other image dimensions can be taken into account (1D, 3D, ..., etc.) [9,10], depending on what kind of imaging systems is used to create digital images. Each A(m, n) element of the array A corresponds to a pixelwhich describes some properties of the image. We can use many shades of grey typically 16 or 256 to represent the pixels. However, grey scanning requires larger amounts of memory. In spite of a greyscale images are simple and have less information in comparison to colour images. It is possible to construct all visible colours by combining the three primary colours: red, green and blue (RGB colour image).

A. Image Matching Algorithms and related definitions

Theimage matching algorithms for the compared images or ROIs regarding the accuracy can be evaluated by the similarity degree, therefore we give the following definition that is needed for the matching problem.

*Def*inition 2.1.

Let be given matrix A1 representing a template image I1 and matrix A2 representing a target image I2. For images I1 and I2 the following three cases should be considered:

1) Images are the same (A1=A2) if and only if similarity criteria SC (A1, A2)=1.

2) Images I1 and I2 are similar if and only if $\Delta \leq SC(A1, A2) < 1$.

3) Images I1 and I2 are different if and only if $0 \leq SC(A1, A2) < \Delta$.

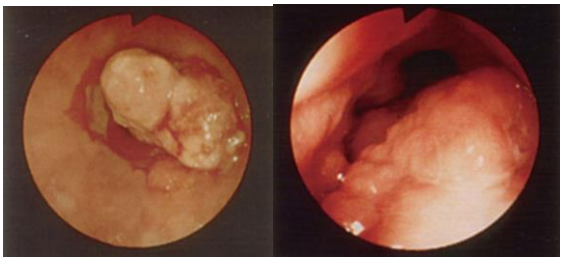
Similarity criteria SC and threshold Δ (SC ε < 0, 1> and Δ ε < 0, 1>) can be chosen arbitrarily for each class of matching algorithms.

In case of pixel to pixel comparison [11,12], we can define similarity criterion SC(Ik, Ik+1) as the following formula:

$$MS = SC\left(I^k, I^{k+1}\right) = \sqrt{\sum_{i=1}^{M}\sum_{j=1}^{N}\left[A^k(i,j) - A^{k+l}(i,j)\right]}^2$$

(2.1)

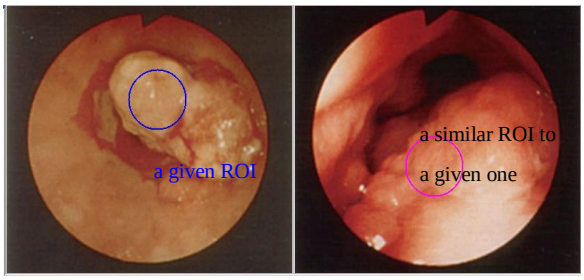
where $A^x(i, j)$ is the pixel digital value for x^{th} image, it can be referred either to the whole image or to its ROI (see Fig. 2, 3). In many cases the similarity degree MS is higher for ROI than for the whole images. In case of ROI the similarity criterion should be suitable modified (i.e. proper pixels are only compared).



a b

similarity degree = 0.8394

Fig. 2. Example of matching two images, a) template, b) target



a b

similarity degree = 0.8718

Fig. 3. Example of matching two ROI, a) template, b) target

B. Image Matching Algorithms

To solve the matching problem we propose four algorithm[13], systematic (lexicographical) searching, gradient searching, random searching and simulated annealing searching, that define the methods of searching the ROI in the target image that is best match the one specified in the template image. In systematic searching algorithm an initial location of the ROI is assumed to be on the left top corner of the target image, the the center of the ROI is moved from left to right and up to down in the target image with specific step of pixels, for each location of the ROI the relative similarity degree is calculated, at the end the location with the best similarity degree is pointed out as the best solution. Gradient searching algorithm can assume a random choice of the initial location of the ROI in the target image, next we calculate the step and the direction of the ROI movement, to find the best matching location we decrease the step twice in each iteration that returned the optimal location, and from that location we repeat this process of searching we get the best matching location. In random searching we determine only the number of iterations and every iteration the location of the ROI is randomly selected, after such process the optimal location with the best matching similarity of the ROI in the target image is returned. Finally in simulated annealing searching algorithm also the initial location of the ROI in the target image is selected randomly with a given number of iterations and with high starting temperature which is reduced in each iteration according to the annealing scheme, the location of the ROI in each location then is changed with probability determined by the generation function and the similarity degree is calculated for the new location with probability determined by the acceptance. After reaching the maximum iteration, we choose the optimal solution the found solutions.

III. Experimental Results and Discussion

Experiments are carried out for endoscopic and the obtained results discussed in order to evaluate the different searching procedures, that helps choosing the best procedure of matching the whole endoscopic image, as well as ROI searching. The representative benchmarks for illustration of our considerations is presented in table 1., where there are different endoscopic images (size of 800 x 720 pixels), chosen from endoscopic atlas [14] among 1500 available images. We concentrate on stomach diseases, and include five images (from 1 to 5), corresponding to healthy patients, next nine images (from 6 to 14) contains some changes referring to typical (representative) stomach diseases. The last five images from 15 -19 represent similar changes regarding to appearance.

TABLE I. Detailed Description of the Endoscopic Images Belonging to the Test Benchmark

The number of image in Fig. 5.1	Its atlas number	Description of changes in the endoscopic images
1	1_24	Formix fundus, healthy
2	1_25	Corpus venriculi, healthy
3	1_26	Corpus ventriculi, healthy
4	1_27	Antrum, anqulus, healthy
5	1_32	Antrum, healthy
6	6_15_a	Ventriculus, cancer
7	6_18_a	Ventriculus, cancer
8	6_18_b	Ventriculus, cancer
9	6_18_c	Ventriculus, cancer
10	6_23	Ventriculus, cancer
11	6_25_b	Ventriculus, cancer
12	6_5	Corpus, cancer
13	6_8_a	Corpus, cancer
14	6_8_b	Corpus, cancer
15	6_27	Fundus ventriculus, cancer
16	6_30_b	Fundus ventriculus, cancer
17	6_32_c	Corpus ventriculus, cancer
18	6_3_a	Corpus, antrum, cancer
19	6_6	Cardia, cancer

A. Image and ROI matching strategies

We consider two kinds of image matching algorithms. The first one oriented on the whole image matching where we compare corresponding pixels or chosen image characteristics.

The second one is oriented on finding the region in the target image which matches best the ROI given in the template image. In our experiments we take the list of target images, directly from the atlas, and a subset of these images as the template images. In case of simple matching criteria (MS – formula (2.1), IF - formula [13] the similarity degree for the whole image gives value 1 for the same image. Table 2. confirms such cases.

Let consider ROI - oriented matching for the same set of template and target images. Note that even in case of the same compared images values of similarity degree are not equal to 1. Moreover, we can find higher value of these criteria for different images than for the same images see the first and the column of Table 3. However, it does not mean that this kind of matching is not practically acceptable.

TABLE II. THE SIMILARITY DEGREE VALUES FOR THE WHOLE IMAGE MATCHING

No of template images / No of target images	1	2	7	11	16	17
1	**1.0000**	0.7910	0.6744	0.7353	0.8140	0.6342
2	0.7910	**1.0000**	0.7529	0.7583	0.7792	0.7437
3	0.7715	0.8285	0.7465	0.7512	0.7945	0.7461
4	0.7479	0.8515	0.7985	0.7351	0.7617	0.7591
5	0.8185	0.8385	0.7415	0.7460	0.8062	0.7311
6	0.6928	0.7135	0.7637	0.7617	0.7734	0.7442
7	0.6744	0.7529	**1.0000**	0.7421	0.7211	0.7845
8	0.7142	0.7805	0.7859	0.7695	0.7562	0.7649
9	0.6677	0.7925	0.7453	0.6623	0.6703	0.7463
10	0.7232	0.7870	0.6684	0.7332	0.7623	0.7463
11	0.7353	0.7583	0.7421	**1.0000**	0.7949	0.7133
12	0.7429	0.8167	0.7500	0.7342	0.7790	0.7706
13	0.6685	0.7579	0.7860	0.7338	0.7417	0.8343
14	0.7964	0.8451	0.7704	0.7711	0.8210	0.7728
15	0.8016	0.8232	0.7691	0.7739	0.8049	0.7482
16	0.8140	0.7792	0.7211	0.7949	**1.0000**	0.7044
17	0.6342	0.7437	0.7845	0.7133	0.7044	**1.0000**
18	0.6478	0.7337	0.7873	0.6771	0.6847	0.7252
19	0.7520	0.7841	0.7450	0.7753	0.8010	0.7500

TABLE III. SIMILARITY DEGREE VALUES FOR ROI - ORIENTED MATCHING

No of template images / No of target images	1	2	7	11	16	17
1	**0.9082**	0.8465	0.7406	0.7566	0.7600	0.7739
2	0.9068	**0.9689**	0.8229	0.8755	0.8177	0.8804
3	0.8553	0.7815	0.7982	0.8021	0.7820	0.8159
4	0.8598	0.8828	0.8513	0.8612	0.8516	0.8752
5	0.8468	0.9031	0.7488	0.7721	0.7768	0.7703
6	0.8710	0.8400	0.8681	0.8895	0.8903	0.9046
7	0.8696	0.8289	**0.9514**	0.9101	0.8708	0.9153
8	0.8799	0.8163	0.8751	0.9002	0.8893	0.9282
9	0.8860	0.8252	0.9297	0.8807	0.8812	0.9041
10	0.8837	0.8299	0.8703	0.8839	0.8715	0.9004
11	**0.9089**	0.8086	0.8616	**0.9262**	0.8830	**0.9401**
12	0.8619	0.8337	0.8878	0.9097	0.8907	0.9359
13	0.8549	0.8253	0.8804	0.8978	0.8852	0.9104
14	0.8671	0.8455	0.8168	0.8849	0.8737	0.8566
15	0.8500	0.8227	0.7825	0.8117	0.8021	0.8238
16	0.8661	0.8164	0.8480	0.8614	**0.9374**	0.8698
17	0.8463	0.8213	0.9023	0.9033	0.8900	**0.9375**
18	0.8657	0.8073	0.9195	0.9183	0.8819	0.9263
19	0.8954	0.8276	0.8842	0.8882	0.8559	0.8989

B. Evaluation of searching procedures

We consider four sequential procedures: systematic, random, gradient and simulation annealing defined in [13]. They operate only on pairs of target/template images where ROI's are determined by experts. In our experiments we assume that the target image is the same as the template one, but without ROI. We made many such experiments, but representative results are shown in Table 4. and Fig. 4., the best results we obtained for simulation annealing procedure, then for gradient procedure, we also note that random and systematic procedures give nearly the same level of the mean accuracy, however they are a bit a lower than the first two procedures.

TABLE IV. THE MEAN IMAGE MATCHING ACCURACY OF SEARCHING PROCEDURES

Searching procedure / No of compared image	Systematic	Gradient	Random	Simulated annealing
4	0.7446	0.6729	0.8416	0.8721
5	0.6297	0.5815	0.5851	0.7792
10	0.7688	0.7648	0.7556	0.8426
16	0.8407	0.8153	0.8276	0.9210
17	0.8265	0.9508	0.8459	0.8556
19	0.7827	0.8910	0.7448	0.7920
Mean value	0.7655	0.7794	0.7648	0.8437

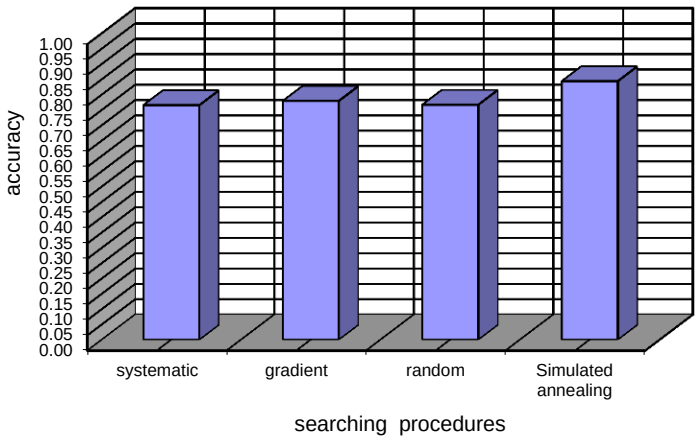


Fig. 5. The mean image matching accuracy of searching procedures

IV. CONCLUSION

In our paper four searching procedures were investigated and analyzed for the endoscopic images. Such images are very heavy for analysis owing to some deformations made during their registrations. Therefore we decide to construct four such algorithms different from each other by use of systematic random, gradient and simulation annealing searching methods. All versions are so flexible that can be tune to improve quality of searching and matching in accordance to specific features of endoscopic images.

For future work further improvement of the algorithms should be done. As well as parallelization of searching procedures will be presented and the version parallel matching algorithms will be described, analyzed and evaluated.

REFERENCES

[1] Elsen P. A. V. D., Pol Evert-Jan D., Viergever M. A.: Medical Image Matching – A Review with Classification. IEEE Trans. on Engineering in Medicine and Biology, March 1993.

[2] ZulaikhaBeevi S. And Sathik M. M. :An Effective Approach for Segmentation of MRI Images: Combining Spatial Information with Fuzzy C-Means Clustering, European Journal of Scientific Research,EuroJournals Publishing, Inc. 2010,http://www.eurojournals.com/ejsr.htm.

[3] Simon R., Krieger D., Znati T., Lofink R., Sclabassi R.J.: Multimedia MedNet - A Medical Collaboration and Consultation System. IEEE Computer, May 1995.

[4] Sacile R., Ruggiero C., Lombardo G., Nicolo G., Wolf B., Petersen R.I.: Collaborative Diagnosis over the Internet: A Working Experience. IEEE Internet Computing, November 1999.

[5] Abrado A., Cassini L.: Embedded Java in a Web-Based Teleradiology System. IEEE Internet Computing, May 1998.

[6] Chanda B. And Majumder D. D.: Digital Image Processing and Analysis, Prentice Hall, 2003.

[7] Gonzalez C. and Woods R. E. :. Digital Image Processing, Second Edition, Prentice Hall, 2002.

[8] Umbaugh S. E. Computer Vision and Image Processing: A Practical Approach Using CVIP tools, Prentice Hall, 1998.

[9] Anil K. J.: Fundamentals of Digital Image Processing. Prentice-Hall International, 1989.

[10] Johnson A., Hebert M.:Using spin images for efficientobject recognition in cluttered 3D scenes,. IEEE Trans. On PAMI , 21(5):433-449, 1999.

[11] Grauman K. and Darrell T. :Efficient Image Matching with Distributions of Local Invariant Features, Proceedings of the IEEE conference on computer Vision and Pattern Recognition (CVPR), June 2005.

[12] Manimala S. and Hemachandran K, "Image Retrieval-Based on Color Histogram and Performance Evaluation of similarity Measurement", Assam University Journal of science & Technology, Vol. 8 Number II 94-104, 2011.

[13] Saif J.: Sequential & parallel image matching algorithms for endoscopic recommendations, PHD Thesis WETI2003, Gdansk, April 2004.

[14] Silverstein F.E., Tytgat G.N.J.: Gastrointestinal Endoscopy. Times Mirror Int. Publisher Limited, 1997.

AUTHOR PROFILES

Khaled M. G. Noaman received the PhD degree in Artificial Intelligent (artificial neural networks) from Wroclaw University of Technology, Wroclaw, Poland, in 1999. In July 2006 he got promoted to associate professor at the Faculty of Computer Science and Information Technology, Sana'a University, Sana'a, Yemen. From 2012 till now he is working in the department of Distance Learning, Deanship of E-Learning and Distance Learning, Jazan University, Jazan, Kingdom of Saudi Arabia.

Jamil A. M. Saif received the Msc degree in telecommunication engineering, from the Department of telecommunication, Faculty of Automatic Control, Electronics and Computer Science , Silesian University of Technology, Gliwice, Poland, in 1993, the PhD degree in Computer Science from the Department of Computer Architecture, Faculty of Electronics, telecommunications and Informatics, Gdansk University of Technology, Gdansk, Poland, in 2004. In May 2012 he got promoted to associate professor at the Faculty of Computer Science and Engineering, Hodeidah University, Hodeidah, Yemen. From 2013 till now he is working in the Department of Information Systems, Community College of Bisha , King Khalid University, Kingdom of Saudi Arabia.

kEFCM: kNN-Based Dynamic Evolving Fuzzy Clustering Method

Shubair Abdulla
Depart. of Instructional & Learning Technologies
Education Colloege
Sultan Qaboos University
Muscat, Oman

Amer Al-Nassiri
IT College
Ajman University of Science and Technology
Fujairah - Campus
United Arab Emirates, Fujairah

***Abstract*—Despite the recent emergence of research, creating an evolving fuzzy clustering method that intelligently copes with huge amount of data streams in the present high-speed networks involves a lot of difficulties. Several efforts have been devoted to enhance traditional clustering techniques into on-line evolving fuzzy able to learn and develop continuously. In line with these efforts, we propose kEFCM, kNN-based evolving fuzzy clustering method. kEFCM overcomes the problems of computational cost, dynamic fuzzy evolving, and clustering complexity of traditional kNN. It employs the least-squares method in determining the cluster center and influential area, as well as the Euclidean distance in identifying the membership degree. It enhances the traditional kNN algorithm by involving only cluster centers in making classification decisions and evolving on-line the clusters when a new data arrives. For evaluation purpose, the experimental results on a collection of benchmark datasets are compared against other well-known clustering methods. The evaluation results approve a good competitive level of kEFCM.**

Keywords—Evolving; Fuzzy Logic; Clustering; k-NN

I. INTRODUCTION

Clustering Analysis is broadly applied successfully in many research areas such as market research, pattern recognition, data analysis, image processing, and document categorization [1] [2] [3] [4]. Clustering aims at describing data by defining set of clusters, which are naturally circles, based on similarities. The approach of finding approximate centroids is commonly used to form the clusters. A cluster centroid is used to determine the cluster location, and later, the system will tell to which cluster a group of input vector belongs by measuring the similarity in predefined features. Forming the clusters also involves determining the influential area of the clusters, which is equal to the radius.

In clustering, there are two crucial terms: fuzzy and evolving. The fuzzy term refers to the overlapping in clusters that is each element in a dataset belongs to one or more cluster in a degree. The cluster belongingness, called fuzzy membership (μ_{ij}), is used to discover the relation between the data element and disclosed clusters. The Euclidean distance is employed commonly to obtain the fuzzy membership values of elements in different clusters, i.e. distance between data point and cluster center. Technically, the evolving term means ability of the system to dynamically updating the clusters, adjusting the clusters centers and/or radius, to accommodate new unseen data when presented.

Beside their ability of analyzing data and making decisions based on acquired intelligence, the evolving clustering methods play an essential role in fuzzy rule-based systems (FRBS) and neuro-fuzzy systems (NFS) which are intelligent systems able to learn and develop continuously in order to enhance their performance. Over the last decade, the evolving clustering methods has boosted the emergence of these systems [5].

Designing an evolving fuzzy clustering algorithm involves a lot of difficulties. In the present high-speed networks, the huge amount of data streams, such as IP flows and network payloads, calls for on-line, fast, non-iterative evolving methods. Dealing efficiently with huge amount of multi-dimensional data items can be problematic because of clustering complexity and computational cost. The algorithm has to perform an incremental learning paradigm that is carried out to update the knowledgebase whenever new data emerges. Moreover, it has to efficiently manage previously seen training data to accommodate new data, and that needs an efficient memory management mechanism.

Unfortunately, most of the data clustering techniques such as K-means [6], Fuzzy C-means, Mountain clustering, and Subtractive clustering [7] lacks these capabilities.

Recently, the issue of creating evolving fuzzy clustering approaches to obtain the best fit of a dataset has been the subject of several research efforts. The research trends may be broadly divided into two directions: (i) to invent new techniques; (ii) to enhance traditional clustering techniques.

The k-Nearest Neighbors (kNN) clustering method [8] is among the clustering techniques in which development has seen attempts. kNN is one of the most simple machine learning methods. It can be used as a baseline for large developmental expansions. It has been selected as one of the top 10 data mining algorithms [9]. However, despite these pros, it has some cons: (i) it is computationally expensive; (ii) it requires large memory; (iii) it does not have ability to learn which data are most important.

In line with the trends that seek to enhance traditional clustering techniques, we present an enhanced version of the kNN algorithm, kNN-based Evolving Fuzzy Clustering Method, kEFCM for short. It is worth mentioning that kEFCM is introduced as a preprocessor for the neural fuzzy inference model [10]. The problems of designing an evolving fuzzy clustering method are addressed through many enhancements

to the original kNN approach such as: reducing the complexity of computation, on-line clustering, and fuzzy evolving. To reduce the computational expense, kEFCM considers the cluster centers only in making classification decisions. The knowledgebase evolving is carried out simply by assigning the coordinates of a new coming example to a new cluster center, and the radius will be the arithmetic mean of all radiuses, in case the example does not belong to any cluster. Neither thresholds nor constraints have been used in the on-line phase.

The rest of this paper is organized as follows. In Section II, we discuss related work, and in Section III, we review the kNN algorithm. The kEFCM approach is explained in Section IV, and we report experiments on real dataset in Section V. Finally, Section VI concludes and indicates the directions for future work.

II. Related Work

Reviewing the literature yields a plenty of clustering approaches. Comprehensive surveys have been published on clustering such as Jiang et al. [11] Xu and Wunsch II [12], and Hruschka et al. [13]. Since they are the main subjects of this paper, we limited our revision to the approaches of evolving fuzzy clustering and to those approaches that are devoted to enhance the kNN clustering method.

The First attempts of data fuzzy clustering could date back to the last century. However, it is still an open problem especially in the present, vast amounts of online information exchange. k-Means clustering [14] is based on finding data clusters such that an objective function of distance (Euclidean distance in most cases) measure is minimized. This algorithm in non-fuzziness and does not solve the overlapping issue. It gives either 1 when a data belongs to a cluster or 0 otherwise. The fuzzy c-means (FCM) is the most popular fuzzy clustering algorithm that also uses an objective function while clustering the data. A given data may belong to several clusters in different digress identified by membership value from 0-1. Since it has a number of drawbacks such as high time requirements, noise, and difficulty in identifying the initial clusters [15], some developments have been suggested. One of these developments is the Possibilistic FCM (PFCM) [16] which is an attempt to solve the noise sensitivity defect of the FCM [13]. The Multi-Kernel Fuzzy Clustering (MKFC) [17] is another attempt to develop the FCM which addresses the problem of limitation to spherical clusters. It incorporates multiple kernels and automatically adjusts the kernel weights to make the system immune to ineffectiveness kernels and irrelevant features.

In 2002, Kasabov and Song [3] introduced the Evolving Fuzzy Clustering Method (ECM), which is considered as first evolving on-line clustering method [18]. ECM operates in two phases: off-line and on-line. In off-line phase, it estimates dynamically the number of clusters in a one-pass algorithm. The number of clusters depends on a threshold value, Dthr, which has to be tuned initially. The Dthr is used to control the maximum distance between a data and the cluster center. In the on-line phase, when ECM receives a data sample, based on its position in the dataset, ECM either creates a new cluster or updates some existing clusters. The value of Dthr is used to control updating cluster centers, if the radius equals to Dthr, the cluster will not be updated.

Some sophisticated fuzzy clustering methods have emerged over the past few years. For example, the Fuzzy Rule-Based Classifier (FRBC) [13] inherently performs the unsupervised cluster analysis by employing a supervised classification approach. It explores the potential clusters and identifies them by using interpretable fuzzy rules. The actual boundaries are revealed through simultaneous classification of data with the fuzzy rules. The Evolving Local Means (ELM) [19] is another example. It is simple and has the desirable features of density based approached. It uses the concept of non-parametric gradient estimate of a density function. The evolving process is performed based when the density pattern changes.

The research of deriving fuzzy clustering methods from the traditional kNN algorithm was initially motivated by its drawbacks. For example, the authors in [20] present a clustering ensemble algorithm based on kNN. To summarize the ensemble data, the algorithm generates the similarity matrix of data and then it uses hierarchical clustering to get the final clustering. Another example can be seen in [21]. It is a special cluster matching algorithm that establishes correspondence among fuzzy clusters by building a new combination model based on cluster matching and fuzzy majority vote. In [22], a new kNN-based clustering method, called kNNModel, is proposed. The model is similar to kNN, but the k value is automatically determined. A data model is built by extracting a set of representatives of the training data. The representatives whose size is far less than the whole training data are involved in making classification decision. The kNNModel is enhanced in [23] by developing a cluster-based training algorithm to learn the optimized set of representations.

In some sense, the performance of the reviewed evolving fuzzy clustering methods is effective. However, we believe that an effective clustering method should possess the features: (1) fuzzy clustering, (2) dynamic evolving, (3) low computational cost, and (4) little efforts for prior tuning. The demand of such method has not been yet achieved. For example, although the ECM is on-line evolving fuzzy clustering, its performance relies on prior precise tuning of Dthr parameter. With respect to the methods that aim at developing kNN algorithm, unfortunately, the dynamic evolving is still a crucial demand. The kEFCM approach is concern about the dynamically evolving which distinguishes it from the above mentioned development of kNN.

III. kNN: k-Nearest Neighbors Algorithm

In this section, we briefly describe the kNN algorithm. kNN is an instance-based learning algorithm. Although it is most often used for classification, it also can be used in estimation and prediction. Given a set of training data, a new data may be classified simply by comparing it to the most similar data in the training dataset. The process of building kNN classifier involves identifying k value, the number of the most similar classes to be considered in the training dataset. The process involves also measuring the similarity based on defining the distance function. The most commonly used distance function in Euclidean distance:

$$d_{Euclidean}(X,Y) = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2} \ldots\ldots \quad (1)$$

Suppose that we are interested in classifying the type of network packet captured by a traffic collector system based on certain characteristics, such as the payload size and the destination port#. For a sample of 200 packets, Figure 1 shows a scatter plot of the packet size against the destination port#. The type of the points symbolizes a particular network packet class. Circle points indicate A class; diamond points indicate B class; square points indicate C class.

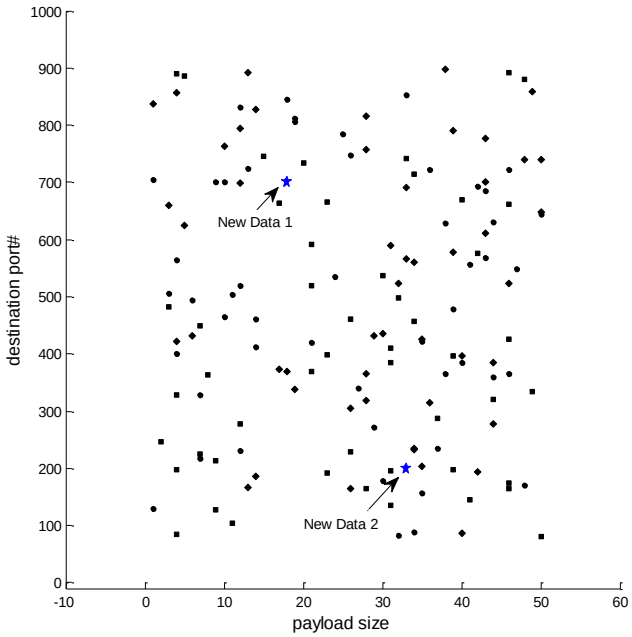


Fig. 1. Scatter plot of payload size against destination port#

Now suppose that there are two new network packets (indicated as stars in the Figure) without classification and would like to classify them based on other packets with similar attributes. New data 1 is composed of 18 bytes and directed to port# 700. Since the packet attributes place it into a section where six packets of the nearest packets belong to C class (square), we would thereby classify it as C easily.

Regarding the new packet 2, which is 33 bytes directed to port# 200, suppose k=1 so that any new data would be classified according to whichever one point it closest to. In this case the packet would be classified into B class since that the closest packet on the scatter plot belongs to B class (diamond point). Suppose we now set k=2 so that the new packet 2 would be classified according to the classification of 2 packets closest to it. One of these packets belongs to C class (square) and one belongs to B class (diamond). The kNN classifier cannot decide between these two classifications. The voting is helpless here since there is one vote for each of two classes. The voting will not help either for k=3 in case of the three nearest packets belong to three different classes.

After determining which training data are most similar to the new unseen data, we need to establish a combination function for classification decision. A combination function could be unweighted voting (each neighbor has one vote) or weighted vote (closer neighbors have larger vote). In either case, this function is computationally expensive.

The above example has shown that the number of nearest neighbors, k, is considered as one of the most influential factors in the accuracy of the classification. The value of k must be set carefully, small value may maximize the probability of misclassification, and large value may make the k nearest packets distant from the right class. The obvious best solution is to employ a cross-validation procedure which is done by trying various values of k with different randomly selected training datasets and determining precisely the k value that minimizes the classification error.

IV. kEFCM: kNN-Based Evolving Fuzzy Clustering Method

The kEFCM runs in two phases: off-line and on-line phase. During the off-line phase, kEFCM partitions the input space into clusters, while in the on-line phase; kEFCM classifies new coming data and updates dynamically the clusters for the purpose of evolving.

A. Off-line Clustering Phase

In the off-line phase, kEFCM applies fast, optimized technique for clustering dataset points. Figure 2 presents a high-level overview of kEFCM in the off-line clustering process.

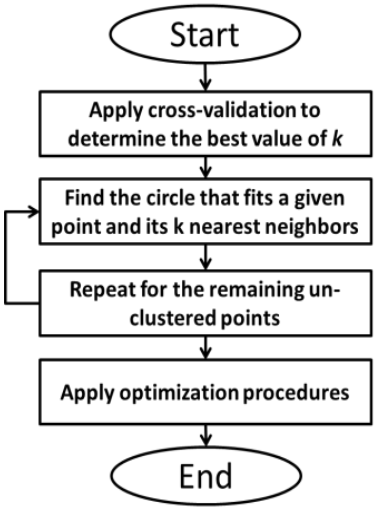


Fig. 2. kEFCM Off-line Clustering Phase

The process of off-line clustering starts by taking the first sample of the dataset (x_1, y_1) and finds its k-nearest points (X_i, Y_i) using the Euclidean distance (eq. 1) where $i = 1,2,\ldots,k+1$. Then the least squares method (LMS) is used to find the equation of the circle that best fits the points (X_i, Y_i) by calculating the center and the radius. A linearized model of the circle equation is needed to determine the values of center (a, b) and radius (r):

$$(x_i - a)^2 + (y_i - b)^2 = r^2 \ldots\ldots\ldots\ldots (2)$$

The linearized model of this equation:

$$x_i^2 - 2ax_i + a^2 + y_i^2 - 2by_i + b^2 = r^2$$

$$x_i^2 + y_i^2 = 2ax_i + 2by_i + r^2 - a^2 - b^2$$

$$x_i^2 + y_i^2 = Ax_i + By_i + C \ldots\ldots\ldots\ldots (3)$$

Equation (3) is now linear with three undetermined coefficients, A, B, and C. In this case, the matrices are used to solve the least squares problem:

$$\begin{bmatrix} A \\ B \\ C \end{bmatrix} = \begin{bmatrix} \sum x_i^2 & \sum x_i y_i & \sum x_i \\ \sum x_i y_i & \sum y_i^2 & \sum y_i \\ \sum x_i & \sum y_i & k+1 \end{bmatrix}^{-1} \begin{bmatrix} \sum x_i (x_i^2 + y_i^2) \\ \sum y_i (x_i^2 + y_i^2) \\ \sum x_i^2 + y_i^2 \end{bmatrix} \ldots\ldots\ldots\ldots (4)$$

After having values for A, B, and C, the circle is simply determined by calculating its center (C_x, C_y) and radius (r):

$$C_x = \frac{A}{2} \ldots\ldots\ldots\ldots (5)$$

$$C_y = \frac{B}{2} \ldots\ldots\ldots\ldots (6)$$

$$r = \frac{\left(\sqrt{4c + A^2 + B^2}\right)}{2} \ldots\ldots\ldots\ldots (7)$$

This process is repeated on the remaining data points. As this step may create unwanted overlapped clusters, the next step applies an optimization procedure that handles two constraints: (1) The number of clusters that contain small cluster(s) is equal to 0; (2) The number of clusters that include points less than k is equal to 0. These constraints are represented mathematically by two functions: probability of inclusion P(I) and probability of violation P(V) respectively:

- Probability of inclusion *P(I)*

$$P(I) = \frac{\sum_{i=1}^{n} \sum_{j=1}^{n} f(C_i, C_j)}{n(n-1)/2} = 0 \ \forall\, i \neq j \ldots\ldots\ldots\ldots (8)$$

Where:

$f(C_i, C_j)$: the inclusion function:

$$f(C_i, C_j) = \begin{cases} 1 & C_j \subset C_i \\ 0 & C_j \not\subset C_i \end{cases}$$

C_i, C_j *:* any cluster, n: # of clusters

- Probability of violation P(V)

$$P(V) = \frac{\sum_{i=1}^{n} f(C_i)}{n} = 0 \ \ldots\ldots\ldots\ldots (9)$$

Where:

$f(C_i)$ *:* violation function,

$$f(C_i) = \begin{cases} 1 & C_i < k \\ 0 & otherwise \end{cases}$$

C_i*:* any cluster, n: # of clusters, k: # of *nearest neighbors* $k < n$

B. Algorithm of kEFCM Off-line Phase

The kEFCM off-line clustering algorithm is given below as pseudo code:

INITIALIZATION:
N: No of samples,
n <= N : No of prototype samples,
k: No of nearest neighbors.
BEGIN %Off-line phase
Step 1: Take a data sample and find its k-nearest samples by using Euclidean distance.
Step 2: By using equations (4), (5), (6), and (7), determine the cluster that fits the sample and its k-nearest samples.
Step 3: Repeat steps 1 and 2 for remaining samples.
Step 4: Find the probability of inclusion P(I) and probability of violation P(V) for the partitioning by using equations (8) and (9).
Step 5: If P(I)=0 and P(V)=0 then STOP
Step 6: Else, remove all inclusion:
$\forall f(C_i, C_j) = 1$
Set $C_i = C_i \cup C_j$
And adjust the # of violated neighbors to $\geq K$
Step 7: Go to Step4
END %Off-line phase

Figures 3-5 explain graphically four cases that possibly happen during the off-line clustering phase, assuming that the total number of staring points is 19 and the optimum value of k is 3.

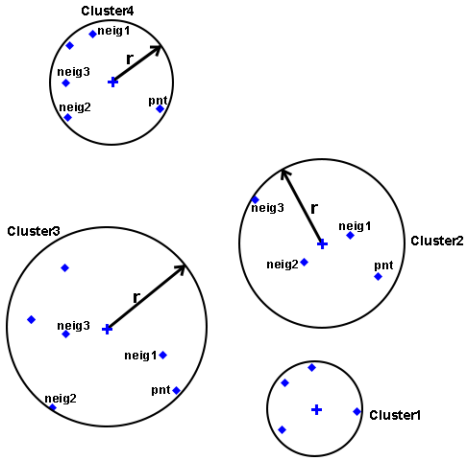


Fig. 3. Case (A): this case is normal case. In cluster 3, the point received by kEFCM is "pnt" and its three neighbors (nieg1, nieg2, and nieg3) have formed a valid cluster. Two points (unlabeled points) will be included in the cluster as they are placed within the cluster influence range

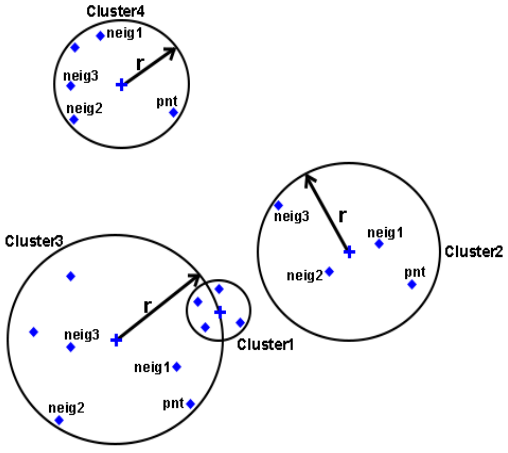


Fig. 4. Case (B): despite that cluster 1 overlaps cluster 3, no optimization is need as no cluster contains small clusters (P(I)=0)

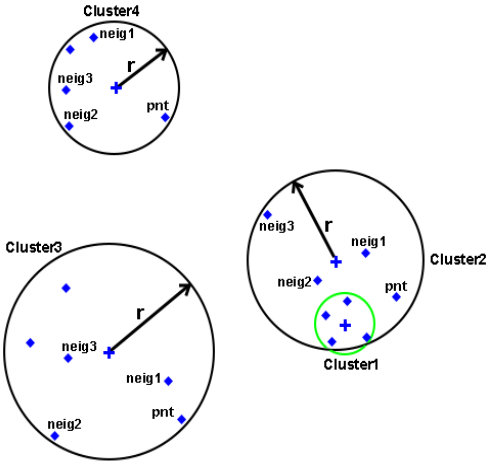


Fig. 5. Case (C): this case represents inclusion violation, cluster 2 contains cluster 1 (P(I)>0)

C. On-line Evolving Phase

This phase of kEFCM classifies new coming data and evolves the clusters dynamically. Figure 6 presents a high-level overview of the on-line process.

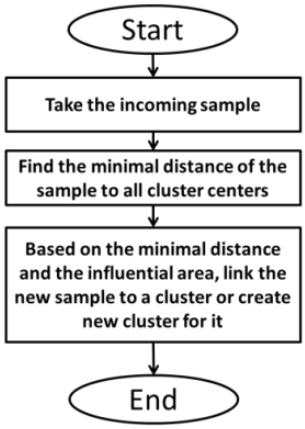


Fig. 6. kEFCM On-line Evolving Phase

The process of evolving operates on the partitioning model that resulted from the off-line phase. Whenever a new data example x is presented to the system, kEFCM updates the clusters according to the position of x. To reduce the complexity inherited from the kNN method, kEFCM computes the distance of x to the cluster centers only instead of computing the distance to all points. If x lies inside the influential range of a cluster, then kEFCM attaches it to the cluster and outputs the cluster class. Otherwise, a new cluster is created by simply assigning the coordinates of x to the center whilst the radius will be the arithmetic mean of all radiuses. The output class in this case is decided based on weighted voting combination function by considering the 3-nearest centers rather than data points.

Any new cluster created dynamically is updated if the number of its points reaches 3, and thereupon, center and radius of the circle that fits the 3 points are calculated by using LSM.

D. Algorithm of kEFCM On-line Phase

The following Algorithm summarizes the evolving process:

INITIALIZATION:
Take the partitioning resulted from off-line phase
BEGIN %On-line phase
Step 1: Take incoming sample x if available
Step 2: find the minimal distance (m) of x to the existing centers. Set the nearest cluster to Cm and its center to Cnm.
Step 3: If m<=Rm, then link the x sample with Cm. output the class of Cm.
a) If the # of samples in Cm =3 then update Cm
Step 4: Else,
a) Create a new cluster Ci+1
b) Center of the cluster Cni+1 = x
c) Calculate the mean of centers M, set Ri+1=M
d) Output the class of the 3-nearest centers by using the weighted voting
Step 5: Go to Step1
END %On-line phase

V. Experiments

This section describes the kEFCM evaluation process. Three sets of experiments were conducted to examine clustering quality, performance, as well as complexity and computational cost. For each set of experiments, we describe the measuring metrics, benchmarking algorithms, and the results of comparison. Before going through these parts, the datasets involved in the evaluation process are described and the results of cross-validation technique used to get optimum value of k are presented.

A. Dataset Used

To assess the quality of clustering of kEFCM, 6 datasets are used in the experiments, 1 forecasting dataset, the gas-furnace [24] and 5 classification datasets selected from KEEL Dataset Repository [25] and UCI Machine Learning Repository [26]. Table 1 summarizes the features and classes of these datasets.

As discussed in Section III, the choice of k is critical. To estimate the value of k accurately, the N-fold cross-validation technique is used. This technique involves setting aside some part of dataset elements for training and the rest for testing. The 10-fold cross-validation has been adopted throughout the experiments. First, the dataset is split into 10 folds. Then, the kEFCM is trained with 9/10 of the dataset, while the reminder 1/10, randomly selected one fold, is used for testing. Five values for k have been suggested: 3, 5, 7, 11, and 13.

TABLE I. Datasets Used for the kEFCM Evaluation

Dataset	Features	Classes	Samples
Gas-furnace	2	-	296
Iris	4	3	150
Glass	9	6	214
Ecoli	7	8	336
Balance Scale	4	3	625
Pima Indian Diabetes (PID)	8	2	768
Heberman Survival (HS)	3	2	306
Relation Banana (RB)	2	2	5292

Eventually, the k value that performed at the highest level of accuracy has been adopted. For the gas-furnace dataset, the k=13 is assigned manually as it has no classes. Table 2 summarizes the results obtained for each value over the dataset used.

TABLE II. N-FOLDS CROSS-VALIDATION RESULTS (%)

Dataset	k-values				
	3	5	7	11	13
Iris	82.9	83.1	**84.1**	84.0	81.8
Glass	88.1	88.1	87.9	88.2	**88.5**
Ecoli	92.7	93.1	94.2	**94.5**	91.8
Balance Scale	83.1	83.5	82.9	83.3	**83.6**
PID	96.4	97.2	97.0	**98.5**	95.8
HS	91.5	92.2	93.6	**94.2**	90.9
RB	61.1	60.5	62.3	62.3	**62.8**

B. *Clustering Quality*

Two parameters were taken as criteria in the comparative analysis:

- **MaxD**: the maximum distance between a point and its cluster center.
- **Cluster Purity:** The quality of cluster:

$$purity = \frac{\sum_{i=1}^{C} \frac{N_i^d}{N_i}}{C} \times 100\% \dots\dots\dots\dots (10)$$

Where C is total clusters, N_i^d is the number of members of the majority class in clusters i, and N_i is the total number of members in cluster i.

The clustering results were compared with those resulted by ECM on-line and off-line. Three experiments were conducted to examine the quality of clustering. In the first experiment, by setting the k value to 13, the kEFCM was employed to cluster the gas-furnace dataset into 15 clusters,

Figure 7 shows the clusters graphically. The results of the second experiment are show in Figure 8 which compares the MaxD values obtained by kEFCM and other clustering methods. In the third experiment, Figure 9, the kEFCM approach (k=7) was used to partition the Iris dataset. By using (eq. 10), the cluster purity is computed for kEFCM and other clustering methods.

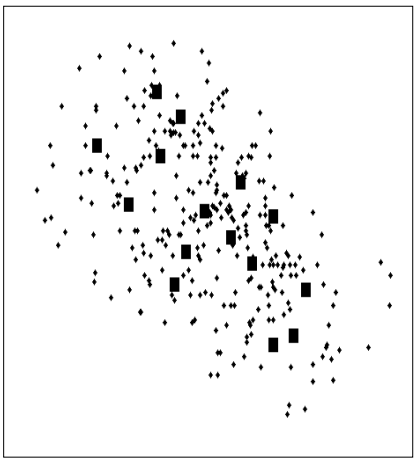


Fig. 7. Clustering gas-furnace dataset into 15 clusters (k=13), ♦: input vector, ■: cluster center

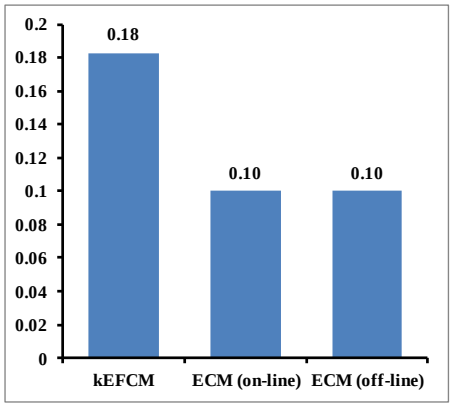


Fig. 8. Comparing the kEFCM against ECM in Terms of MaxD Over Gas-furnace Dataset

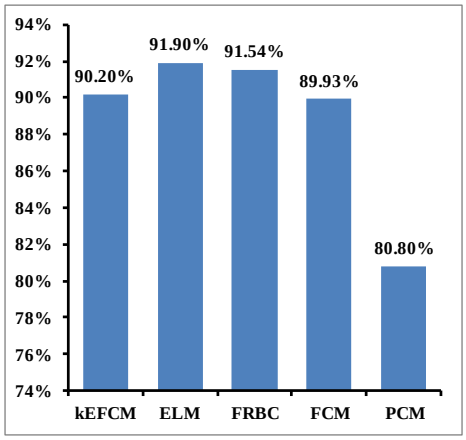


Fig. 9. Comparing the kEFCM against Fuzzy Clustering Methods in Terms of Clusters Purity Over Iris Dataset

TABLE III. COMPARING THE KEFCM AGAINST FUZZY CLUSTERING METHODS IN TERMS OF NMI OVER MULTIPLE DATASETS

Methods	Datasets				Average
	Glass (k=13)	Ecoli (k=11)	Balance Scale (k=13)	PID (k=11)	
kEFCM	0.341	**0.601**	**0.175**	0.116	**0.308**
k-Mean	0.320	0.570	0.121	0.102	0.278
FCM	0.333	0.574	0.118	0.114	0.285
MKFC	**0.355**	0.574	0.120	**0.140**	0.297

TABLE IV. COMPARING THE KEFCM AGAINST FUZZY CLUSTERING METHODS IN TERMS OF ARI OVER MULTIPLE DATASETS

Methods	Datasets				Average
	Glass (k=13)	Ecoli (k=11)	Balance Scale (k=13)	PID (k=11)	
kEFCM	0.177	0.384	**0.139**	0.140	0.210
k-Mean	0.172	0.384	0.129	0.136	0.205
FCM	**0.181**	**0.387**	0.138	**0.143**	**0.212**
MKFC	0.179	0.383	0.135	0.116	0.203

The following points can be concluded from the results:

- Although the perfect value of MaxD was obtained by ECMs, kEFCM achieved very close value, 0.180.
- We computed the standard deviation (stdev) of MaxD for all gas-furnace clusters to check the consistency in the size of clusters. The obtained value 0.3221 shows good consistency.
- Despite the fact that kEFCM is a single distance-based method and may create large number of unstable-mixed-class clusters [27], its ability to remove this kind of clusters is an advantage. kEFCM is equipped with optimization procedure that mainly works against unwanted clusters. It handles two constraints: P(I)=0 and P(V)= 0. The cluster purity reflects this ability, in contrast with ELM and FRBC, kEFCM performed at a comparative value 90.20%, which means that kEFCM produces small rates of unstable-mixed-class clusters. Also, kEFCM outperformed both FCM and PCM clustering methods.

C. kEFCM Performance

We used two common performance metrics to examine the kEFCM in terms of overlapping: Normalized Mutual Information (NMI) and Adjusted Rand Index (ARI) [28]. These metrics compute the level of similarity between clustering system resulted by a method and compare it against the ground truth classes. A higher value means better clustering. There values ranged from 0-1, 1 means perfect match.

Equation 11 is used to compute the NMI of two clustering: C1 clustering resulted and C2 ground truth clustering, of a dataset X of n objects:

$$NMI(C1, C2) = \frac{I(C1, C2)}{\sqrt{H(C1)H(C2)}} \dots\dots\dots\dots (11)$$

Where:

$I(C1, C2)$ the mutual information between *C1* and *C2.*

H(C1) and *H(C2)* the entropy of *C1* and *C2*.

Regarding the ARI, the following equation is used:

$$ARI = \frac{a - \frac{(a+b)(a+c)}{a+b+c+d}}{\frac{(a+b)+(a+c)}{2} - \frac{(a+b)(a+c)}{a+b+c+d}} \dots\dots\dots\dots (12)$$

Where:

a = # pairs of data that are in the same class in $C1_i$ and same cluster in $C2_j$,

b = # pairs of data that are in the same cluster in $C2_j$, but not the same class in $C1_i$,

c = # pairs of data in the same class in $C1_i$, but not the same cluster in $C2_j$, and

d = # pairs of data that are not in the same cluster in $C2_j$ nor class in $C1_i$.

In Tables 3 and 4, we present the NMI and ARI values over multiple datasets for kEFCM and different methods. The last column (Average) of Table 3 displays the average of NMI value for each method over 4 datasets. kEFCM has the best average NMI over all methods. For each individual dataset, kEFCM outperforms all clustering methods in two datasets (Ecoli and Balance Scale), while in the other two datasets (Glass and PID), it is only outperformed by MKFC to be ranked as the second best method. Table 4 presents the results in terms of ARI. The results are slightly changed in contrast to NMI. The kEFCM is the second best in terms of average ARI. It is ranked first for only one dataset (Balance Scale) and ranked second for the remaining datasets. However, despite that, kEFCM, in overall results, has yielded a comparable stable performance.

D. Computational Time & Clustering Complexity

As discussed in Section IV, initially, kEFCM takes a data points and searches for a cluster that best fits the point with its k nearest data points. Then, it loops through the rest of the un-clustered points, each iteration of the loop repeats the same process. To prevent the unwanted overlapping clusters, kEFCM applies equations 8 and 9.

This set of experiments is devoted to examine the computational time along with complexity of the cluster resulting. For the purposes of comparing, we chose FRBC and FCM clustering methods. The results on Iris, Glass, and Ecoli datasets, which are appeared in related research papers, are

compared with those obtained by kEFCM in Table 5. Although the kEFCM consumes more computational time than FRBC and FCM, it is obvious that there is a few timing differences. Add to this, the computational time depends directly on the number of samples within the dataset.

TABLE V. COMPUTATIONAL TIME (SEC) OF KEFCM, FRBC, AND FCM

Methods	Datasets		
	Iris (k=13)	Glass (k=7)	Ecoli (k=11)
kEFCM	1.330	1.410	1.472
FRBC	1.000	1.000	1.200
FCM	0.170	0.100	0.100

With respect to the clustering complexity, according to our view, the complexity of clustering means:

1) Creation of a large number of clusters in off-line phase.

2) Generation of clusters is increasing exponentially in on-line phase.

It has been noted that the number of clusters highly depends on k value, the number of nearest neighbors, which means that the number of clusters is subject to control by the user. Any value of k gives a highly accurate result has to be adopted, since the accuracy is the most important criterion. However, in general, kEFCM shows adequate stability and constant evolution throughout the testing. Figure 10 illustrates two examples of cluster evolutions on different datasets. In the first example of the Heberman survival dataset (k=11), 6 clusters were created off-line to accommodate 10 samples. In on-line phase, when new 90 samples were introduced, it created 14 new clusters to accommodate them. In the second example, the Relational Banana dataset (k=13), kEFCM created only 5 clusters off-line to accommodate 10 samples, and then it created 24 clusters on-line to accommodate new 90 samples. Despite that kEFCM started in both examples with a big number of clusters, it created a very small number of clusters in on-line phase, which means, also, an effective way in clustering unseen samples dynamically.

VI. CONCLUSION

We have proposed in this paper kEFCM, kNN-based evolving fuzzy clustering method. It is an enhanced version of traditional kNN machine learning. kEFCM approach uses the least-squares method for determining the cluster center and radius. The Euclidean distance is used to reflect the membership of a data point in a cluster. The method performs an optimization procedure that handles two constraints, probability of inclusion P(I)=0 and probability of violation P(V)=0. In on-line phase, kEFCM is able to carry out the incremental learning, which is the core tool of evolving. It reduces the computational time that is inherited from kNN by involving the cluster centers in making classification decision.

The clustering ability of kEFCM was examined by benchmarking a collection of real-world datasets. The results obtained were compared against several well-known clustering methods. The results showed that the kEFCM performs at a good competitive level. The possible future work will turn to deploying kEFCM onto real-world environment, where intuitively, it will perform at the same level of success.

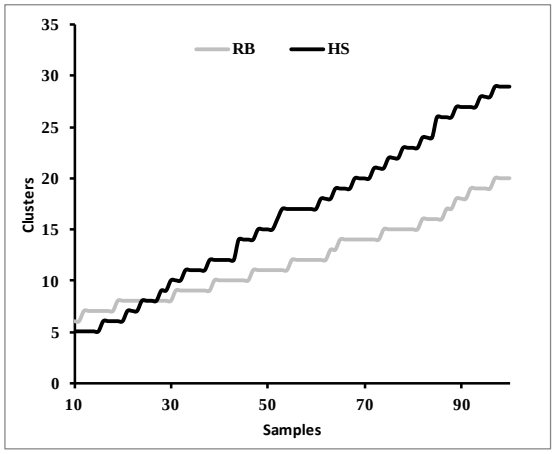


Fig. 10. Cluster Evolutions

REFERENCES

[1] F. Nie and P. Zhang, "Fuzzy Partition and Correlation for Image Segmentation with Differential Evolution," IAENG International Journal of Computer Science, vol. 40, pp. 164-172, 2013.

[2] R. C. D. A. K. Jain, Algorithms for Clustering Data: Prentice Hall, 1988.

[3] G. Mecca, S. Raunich, and A. Pappalardo, "A new algorithm for clustering search results," Data & Knowledge Engineering, vol. 62, pp. 504-522, 2007.

[4] A. K. Abd-Elaal, H. A. Hefny, and A. H. Abd-Elwahab, "Forecasting of Egypt Wheat Imports Using Multivariate Fuzzy Time Series Model Based on Fuzzy Clustering," IAENG International Journal of Computer Science, vol. 40, pp. 230-237, 2013.

[5] N. K. Kasabov and Q. Song, "DENFIS: dynamic evolving neural-fuzzy inference system and its application for time-series prediction," Fuzzy Systems, IEEE Transactions on, vol. 10, pp. 144-154, 2002.

[6] J. M. Keller, M. R. Gray, and J. A. Givens, "A fuzzy k-nearest neighbor algorithm," Systems, Man and Cybernetics, IEEE Transactions on, pp. 580-585, 1985.

[7] P. K. J. a. S. Chattopadhyay, "Comparative Study of Fuzzy k-Nearest Neighbor and Fuzzy C-means Algorithms," International Journal of Computer Applications, vol. 57, p. 10, November 2012.

[8] D. W. Aha, "Editorial," Artificial Intelligence Review, vol. 11, pp. 7-10, 1997.

[9] X. Wu, V. Kumar, J. R. Quinlan, J. Ghosh, Q. Yang, H. Motoda, G. J. McLachlan, A. Ng, B. Liu, and S. Y. Philip, "Top 10 algorithms in data mining," Knowledge and Information Systems, vol. 14, pp. 1-37, 2008.

[10] A. Shubair, S. Ramadass, and A. A. Altyeb, "kENFIS: kNN-based evolving neuro-fuzzy inference system for computer worms detection," Journal of Intelligent and Fuzzy Systems.

[11] D. Jiang, C. Tang, and A. Zhang, "Cluster analysis for gene expression data: A survey," Knowledge and Data Engineering, IEEE Transactions on, vol. 16, pp. 1370-1386, 2004.

[12] R. Xu and D. Wunsch, "Survey of clustering algorithms," Neural Networks, IEEE Transactions on, vol. 16, pp. 645-678, 2005.

[13] E. R. Hruschka, R. J. G. B. Campello, A. A. Freitas, and A. P. L. F. De Carvalho, "A survey of evolutionary algorithms for clustering," Systems, Man, and Cybernetics, Part C: Applications and Reviews, IEEE Transactions on, vol. 39, pp. 133-155, 2009.

[14] J. A. Hartigan and M. A. Wong, "Algorithm AS 136: A k-means clustering algorithm," Journal of the Royal Statistical Society. Series C (Applied Statistics), vol. 28, pp. 100-108, 1979.

[15] E. G. Mansoori, "FRBC: A fuzzy rule-based clustering algorithm," Fuzzy Systems, IEEE Transactions on, vol. 19, pp. 960-971, 2011.

[16] N. R. Pal, K. Pal, J. M. Keller, and J. C. Bezdek, "A possibilistic fuzzy c-means clustering algorithm," Fuzzy Systems, IEEE Transactions on, vol. 13, pp. 517-530, 2005.

[17] H.-C. Huang, Y.-Y. Chuang, and C.-S. Chen, "Multiple kernel fuzzy clustering," Fuzzy Systems, IEEE Transactions on, vol. 20, pp. 120-134, 2012.

[18] V. Ravi, E. Srinivas, and N. Kasabov, "On-line evolving fuzzy clustering," in Conference on Computational Intelligence and Multimedia Applications, 2007. International Conference on, 2007, pp. 347-351.

[19] R. Dutta Baruah and P. Angelov, "Evolving local means method for clustering of streaming data," in Fuzzy Systems (FUZZ-IEEE), 2012 IEEE International Conference on, 2012, pp. 1-8.

[20] F. Weng, Q. Jiang, L. Chen, and Z. Hong, "Clustering ensemble based on the fuzzy KNN algorithm," in Software Engineering, Artificial Intelligence, Networking, and Parallel/Distributed Computing, 2007. SNPD 2007. Eighth ACIS International Conference on, 2007, pp. 1001-1006.

[21] C. sheng Li, Y. nan Wang, and H. dong Yang, "Combining Fuzzy partitions Using Fuzzy Majority Vote and KNN," Journal of Computers, vol. 5, pp. 791-798, 2010.

[22] G. Guo, H. Wang, D. Bell, Y. Bi, and K. Greer, "KNN model-based approach in classification," in On The Move to Meaningful Internet Systems 2003: CoopIS, DOA, and ODBASE, ed: Springer, 2003, pp. 986-996.

[23] L. Chen, G. Guo, and S. Wang, "Nearest neighbor classification by partially fuzzy clustering," in Advanced Information Networking and Applications Workshops (WAINA), 2012 26th International Conference on, 2012, pp. 789-794.

[24] J. D. Farmer and J. J. Sidorowich, "Predicting chaotic time series," Physical review letters, vol. 59, p. 845, 1987.

[25] J. Alcalá, A. Fernández, J. Luengo, J. Derrac, S. García, L. Sánchez, and F. Herrera, "KEEL data-mining software tool: Data set repository, integration of algorithms and experimental analysis framework," Journal of Multiple-Valued Logic and Soft Computing, 2010.

[26] A. Asuncion and D. J. Newman, "UCI machine learning repository," ed, 2007.

[27] E. Lughofer, Evolving fuzzy systems-Methodologies, advanced concepts and applications: Springer, 2011.

[28] J. V. de Oliveira and W. Pedrycz, Advances in fuzzy clustering and its applications: Wiley Online Library, 2007.

Constraint on Repair Resources, Optimal Number of Repairers and Optimal Size of a Serviced System

Marin Todinov
Department of Computing and Communication Technologies
Oxford Brookes University
Wheatley, Oxford

***Abstract*—The focus of this paper is the analysis of the constraint on the repair resources caused by breakdowns of components in large systems. The study has been conducted by creating a very efficient discrete-event simulator, based on a min-heap data structure, for determining the probability of constraint on the repair resources.**

In finding the right balance between the number of repairers and salary costs, an exact optimisation algorithm has been proposed for the first time. The algorithm determines the optimal number of repairers which guarantees that the probability of constraint on the repair resources will not exceed an acceptable tolerable level. In addition, an exact optimisation algorithm has been proposed for the first time, for determining the maximum size of the system that can be serviced by a specified number of repairers so that the probability of constraint on the repair resources remains below a specified tolerable level. Unlike heuristic optimisation algorithms, the proposed algorithms are exact and always guarantee optimal solutions.

The presented results are of significant importance to operators of computer networks, production systems, transportation networks, water distribution systems, electrical distribution networks etc. They are a solid basis for management decisions regarding the optimal number of maintenance personnel needed to service the breakdowns in large systems. Increasing the number of repairers beyond the optimal level leads to high salary costs while reducing the number of repairers below the optimal number leads to a poor quality of service.

Keywords—constraint on the repair resources; discrete-event simulation; optimization; repairs; optimal size of a system

I. INTRODUCTION

Complex systems include many components experiencing failures at random times and with random repair durations.

The constraint on the repair resources caused by overlapping repair intervals and insufficient number of repairers is an important factor which decreases the availability and quality of service of computer systems.

A particular pattern of random breakdowns combined with a random duration of the repair has been shown in Fig.1. The breakdown events have been marked by b1,b2,b3,..., while the return from repair events corresponding to the breakdown events have been denoted by r1,r2,r3,....

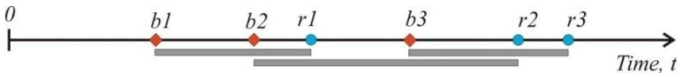


Fig. 1. A specific pattern of random breakdowns and random repair times

Each breakdown engages at least one repairer. In the zones where the repair times overlap, for example zones (*b2*, *r1*) and (*b3*, *r2*), the repairers are engaged in more than one breakdown. This means that several members of the repair team will be simultaneously engaged in fixing breakdowns. The quantity of required repairers depends on the type of breakdown. The return-from-repair events (*r1*,*r2* and *r3* in Fig.1) are associated with restoring the failed devices to a full working state and a release of engaged repairers.

If for a particular breakdown event, the number of remaining repairers is not sufficient to cover the corresponding repair, a constraint on the repair resources occurs, which will result in delays and degraded quality of service of the system.

If the quantity of repairers is more than the optimal quantity, the result will be overstaffing and extra costs. Consequently, there is a delicate balance between the number of repairers and the probability of a delay caused by overlapping repairs. This balance can be found by creating an optimisation model which involves numerous simulations of breakdown-repair histories revealing the probability of a delay caused by insufficient number of available repairers. The maximum tolerable probability of such a delay will be specified in advance, in order to guarantee the required quality of service of the system.

Monte Carlo simulation techniques and discrete-event simulation techniques [1] have become the methods of choice for studying the behavior of complex systems. Monte Carlo simulations related to studying the behavior of the various layers building computer networks already exist. The benefits of simulation techniques applied to analysis of computer systems, modelling and design have been highlighted in [2]. Simulations have also been conducted related to the data traffic carried by computer networks [3,4]. The study of comprehensive review articles on simulation of telecommunication networks [5] shows that, to the best of our knowledge, the constraint on repair resources caused by overlaping repair times *has not yet been studied.* This gap defines the objectives of the present paper:

1) To analyse the constraint on the repair resources caused by component breakdowns in large systems;

2) To determine the optimal number of repairers needed to reduce the probability of a constraint on the repair resources below a specified level;

3) To determine the maximum size of the system that can be serviced by a specified number of repairers, for a specified probability of constraint on the repair resources

Breakdown/failure data related to large systems are vital both in terms of optimal system design and system evaluation. The availability of breakdown data is particularly important to computer networks. According to a number of sources [6,7,10] the negative exponential distribution is an appropriate model for random breakdowns of electronic devices in a given time interval.

If the breakdown density (number of random breakdowns per unit time interval) is denoted by λ, the time to a random breakdown is given by the negative exponential distribution

$$B(t) = 1 - \exp(-\lambda t) \quad (1)$$

where $B(t)$ is the probability that the time to a breakdown will be smaller than or equal to a specified time *t*. The random delay for repair after a random breakdown was modelled by the negative exponential distribution

$$R(t) = 1 - \exp(-t / MTTR) \quad (2)$$

where $R(t)$ is the probability that the repair time will be smaller than a given value *t*, and $MTTR$ is the mean time to repair. The negative exponential distribution (2) has been traditionally used for modelling service times [8]. A recent study [9] has indicated that the assumption that repair times follow the negative exponential distribution practically does not affect the calculated availability of various complex systems.

The breakdown/failure density λ in the negative exponential distribution (1) can be estimated from breakdown/failure data related to the components building the computer system. From *n* recorded times to a breakdown t_1, t_2,...,t_n, the mean time to failure/breakdown $MTTF$ can be estimated [6,7] from

$$MTTF \approx \frac{t_1 + t_2 + ... + t_n}{n} \quad (3)$$

For time to a breakdown following the negative exponential distribution (1), it can be shown that the breakdown/failure frequency λ is related to the mean time to failure (*MTTF*) by the simple relationship [6,7,10]:

$$\lambda = \frac{1}{MTTF} \quad (4)$$

from which the breakdown/failure frequency λ can be determined if the MTTF is available and *vice versa*.

In developing the model of random times to a breakdown for the separate components, breakdown frequencies determined from analysis of past data published in the literature (e.g. [11]) will be used.

A failed device in the computer system comes back in operation after a delay specified by the time it takes for the component to be fixed [10]. As a result, for any component/device building the computer system, each breakdown event is followed by a random delay for repair (Fig.2) in the interval (0, op_int), during which the computer system is operated.

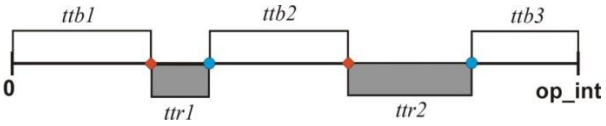


Fig. 2. Random times to breakdown ttb1,ttb2 and ttb3 and random times to repair ttr1,ttr2 for a selected component building the computer system.

Gaining access to failure data is often difficult because failure data are sensitive and not readily disclosed by manufacturers. A number of studies on breakdowns in computer systems already exist ([12, 13,14,15,16,17]. A major drawback in many of these studies ([14], [15], [16], [17]) is that the breakdowns are not attached to specific devices and components. For example, Plank and Elwasif [14] provide time to a breakdown and time to recovery histograms, but these are related to the systems as a whole, not to particular components building the investigated systems. Murphy and Gent [16] discuss system crashes caused by 'software failures', 'hardware failures', 'system management' and 'other reasons', without providing specific data relevant to the separate components building the systems. A similar discussion has been presented by Kalyanakrishnan et al [17], who discuss breakdowns in terms of 'hardware or firmware problems', 'connectivity problems', 'crucial application failures' and problems with a software component'.

The analyses are relevant only to the particular investigated system but not very useful for modelling the behaviour of newly designed systems with similar components, because no sufficient specific failure data were presented about the components building the systems. The research conducted as part of the present study could not identify anonymized failure databases similar to existing databases for the military electronic equipment (for example MIL-HDBK-217F [18]), where electronic components are listed with their breakdown/failure rates. The existence of such databases could help significantly the assessment of new computer systems.

The need for such databases is reinforced by studies [13] indicating that the component failures are not uniformly distributed among the different devices building the computer systems. A small fraction of devices are responsible for the majority of the recorded failures and this circumstance is vital for modelling the expected availability of new systems. Some devices, due to their nature, are subjected to a much higher workload compared to other devices and are more prone to breakdowns (e.g. file servers conducting a large number of I/O operations). Some devices including electro-mechanical parts (disk drives) are much more prone to breakdowns compared to devices which do not include such parts.

Despite these difficulties, failure data related to specific componets in computer networks have been presented by Gill et al [11], Schroeder and Gibson [19,20], Labotitz and Ahuja [21] and Sahoo et al, [13].

In their analysis, Schroeder and Gibson [19] noted that the system size is not a significant factor in the repair time. A set of failure data has been recently collected at a high-performance computing site and was made available online (Failure data [22]). Statistical failure analysis of web server systems has been presented by Fujii and Dohi [23], where failure rate functions characterising web servers can be found.

Valuable failure data sets have been presented by Gill et al [11], collected by computer systems operators employing a ticketing system. The tickets contain important information about when and how the breakdown events have been discovered as well as when they were resolved. In addition, the tickets also contain a description of the cause of the problem, and the specific device at fault. Event logs were collected during one year of operation and two types of breakdowns have been defined: 'link failures' in the case where the connection between two devices was down and 'device failures', when a device was not functioning for routing/forwarding traffic.

II. Determining the Optimal Number of Repairers and the Optimal Size of the Serviced System

A. Determining the optimal number of repairers which guarantee a probability of constraint on the repair resources below a specified tolerable level

Determining analytically the probability of constraint on the repair resources for more than a single available repairer is a complicated task. This task can be simplified by constructing a discrete-event simulator for modelling the failure-repair history. Furthermore, there is an optimal balance between the number of repairers and the probability of constraint on the repair resources. This optimal balance can be found by creating an optimisation model that involves a suitable variation of the number of repairers followed by determining the probability of constraint on the repair resources, until the optimal balance is achieved.

An initial amount of repair resources is assigned and the discrete-event simulator is called to calculate the probability of constraint on the repair resources. Clearly, for zero number of repairers, the probability of constraint on the repair resources will be higher than the target probability α and close to unity (Fig.3). This is because any single component breakdown will create a constraint on the repair resources simply because there will be no available repairers to handle the breakdown.

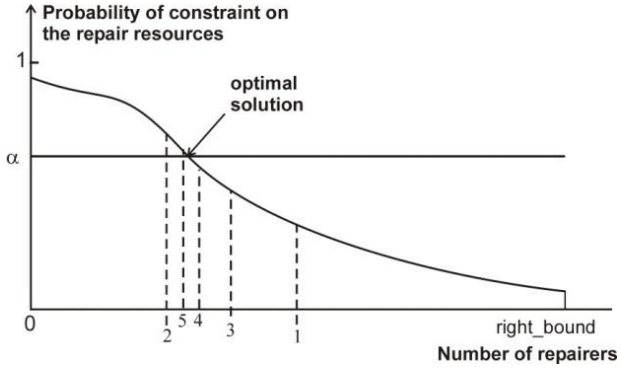


Fig. 3. A procedure for determining the optimal number of repairers for a specified probability α of constraint on the repair resources.

Similarly, for a large number of repairers (number of repairers = right_bound), it is highly likely that there will be an available repairer for every single breakdown and the probaility of runing out of repairers will be smaller than the target probability α . Between these two extremes lies the optimal solution where the number of repairers is just enough to guarantee the target probability α of constraint on the repair resources. This optimal solution can be found by a repeated bisection of the interval (0, right_bound).

The probability of constraint on the repair resources is determined from risk management considerations and is specified as an input paratemer by the manager of the computer system. This probability depends on the criticality of the supplied service. A probability of constraint on the repair resources equal to 20% may be sufficient for non-critical computer systems (e.g. school network; computer LANs) but it is insufficient for critical computer sustems (e.g. server backbone networks, nuclear power plant moniotoring systems, computer networks providing high-speed data transfer for electrical distribution systems; data storage computer systems; secure remote access networks, etc.). For critical computer systems, a large probability of a constraint on the repair resources means a large probability that repairers will not be available when needed, which could be associated with serious consequences. Setting the correct level of the probability of constraint on the repair resources is done by experts, after a careful risk assessement and is not a subject of this study.

B. Determining the maximum size of the system which guarantees a probability of constraint on the repair resources below a specified tolerable level

For a given number of repairers, the maximum size of the computer system which guarantees a probability of constraint on the repair resources below a specified tolerable level can be determined by a similar repeated bisection technique (Fig.4).

This time, the parameter which is varied is the number of components in the computer system while the number of repairers is kept constant.

Clearly, for zero number of components, the probability of constraint on the repair resources is zero. For a large number of components experiencing breakdowns, the probability of constraint on the repair resources will be larger than the specified tolerable level α. The optimal solution is located between these two extremes. It corresponds to a point where the number of components in the system is just enough to guarantee the target probability α of constraint on the repair resources (Fig.4). The optimal solution can be found by a repeated bisection. A serviced system size, significantly smaller than the optimal solution, means that the available repairers are not used efficiently. A serviced system size, significantly larger than the optimal solution, means that there exists an increased probability of constraint on the repair resourecs and a low quality of service.

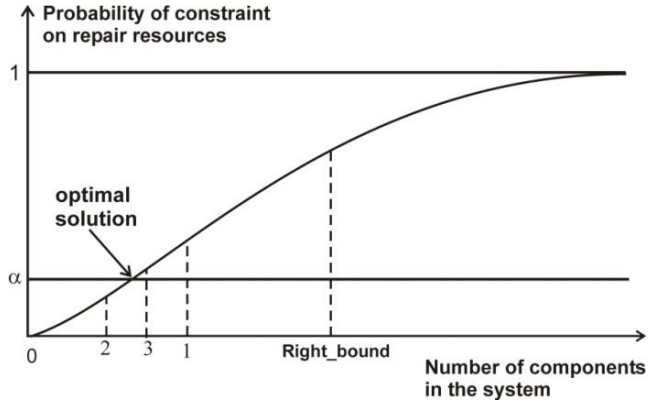


Fig. 4. An optimisation procedure for determining the optimal size of the serviced system

C. An optimization algorithm

The breakdown events mark the engagement of a repairer. The return-from-repair events mark the release of repairers. Each event is represented as a record ‘ev_record’ with four fields:

'time_of_event' - stands for the time of occurence of the event;

'e-type' - stands for the type of event ('breakdown' or 'return from repair');

'req_rep' - stands for the quantity of engaged/released repairers;

'id' - stands for the component index.

The traditional way of implementing discrete-event simulators is by linked lists [1,24]. In this study, the discrete-event simulator has been implemented by using a min-heap (priority queue). The main reason is that the retrieval of the event with the smallest time from a linked list is an operation of time complexity $O(n)$, where *n* is the number of events in the list. In contrast, the retrieval of the event with the smallest time, followed by restoring the min-heap property, is an operation of time complexity $O(\log_2 n)$.

The events are placed in a min-heap which is essentially a binary tree coded in an ordinary array. In the min-heap, the time stamp 'te' of each predecessor node is smaller than each of the time stamps of its successor nodes (te1 < te2; te1< te3; te2<te4; te2<te5; te3<te6; te3<te7; te4<te8; te4<te9).

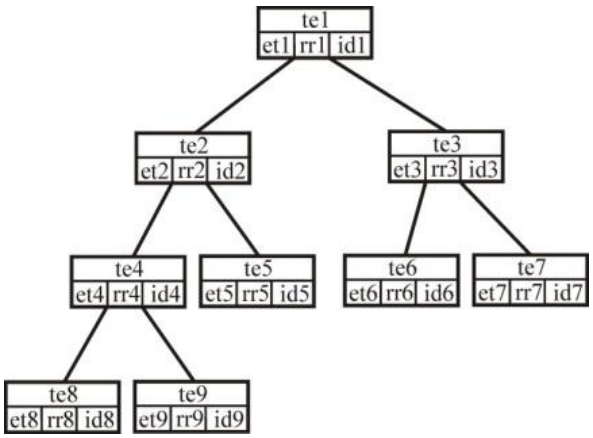


Fig. 5. Min-heap used for implementing the discrete-event simulator.

As a result, inserting an event in a min-heap with *n* events or removing an event, followed by restoring the min-heap property, are both operations of time complexity $O(\log_2 n)$. Standard functions ‘*sift*’ and ‘*buble up*’ can be used to restore the basic property of the min-hip: the descendants of each element of the min-heap must be with a larger time stamp. Detailed discussions regarding the programming implementation of min-heaps can be found in [25].

Algorithm

```
procedure
generate_ttb_and_ret_from_repair(current_time, j);
 {
  // Generates the time to breakdown event and the
    return-from-repair event for component i and places
    these events in the min-heap
 }
procedure remove_event()
   {
     // Removes the event with the smallest time from the top of the
        min-heap
   }
function prob_constraint_rep_res(num_repairers)
{

count=0;   // Counts the number of simulation trials  during which
              a constraint on repairers has occurred

for i=1 to num_simul_trials do
   {
    // initialises the quantity of remaining repair resources
    rem_resources = num_repairers;

       // initialise the current time of the discrete-event simulator
    current_time = 0;

    heapsize=0;   //clears the min-heap
// for all components generates events 'time of a breakdown' and
'time of return from repair' and places  them in the min-heap:
```

```
for j=1 to num_components do
generate_ttb_and_ret_from_repair(current_time,j);
// while there are events in the min-heap and the the current time is
smaller than the operational time do the while loop:
  while (heapsize>0 and current_time < op_int) do
         {
          ev=a[1]; //Take the event on the top of the min-heap
          // Takes the time stamp of the top event in the min-heap
          current_time = ev.time_of_event;
        if (ev.e_type = 'breakdown')  {
            // checks for a constraint on the repair resources
           if (rem_resources < ev.req_rep)
            {
              count=count+1;
              break;
            }
     else {
       // allocates resources for the current breakdown event
       rem_resources = rem_resources - ev.req_rep;

    // removes the first event from the min-heap;
    remove_event();
          }
                                              }
 else { // the event is a return-from- repair event
       rem_resources = rem_resources + ev.req_rep;
        remove_event();
 // generate time of a breakdown event and time of a recovery event
   and place them in the min-heap;
generate_ttf_and_ret_from_repair(current_time,
                                    ev.id);
      }
           }
    }
prob_of_constr = count/num_simul_trials;
return prob_of_constr;
}
// Code of the repeated bisection part
  left=0; right = right_bound;
 while (left + 1 < right) do
{
 mid = (left+right) / 2;
  cur_probability = prob_constraint_rep_res(mid);
    if (cur_probability <=alfa ) right = mid;
      else left = mid;
}
    optim_num_rep = right;
```

The probability of constraint on the repair resources depends on the number of available repairers which is an important parameter of the function ***prob_constraint_rep_res***. This parameter is passed to the function through the variable '`num_repairers`'. At any point of the simulation, the variable '`rem_resources`' shows the current number of remaining repairers. The variable '`current_time`' tracks the time of occurence of the current event. The content of the clock '`current_time`' is compared to the length of the operation interval '`op_int`' and if this is exceeded, the current simulation is terminated. At the start of each simulation trial, an empty min-heap is initialised by making the size of the min-heap zero with the statement '`heapsize=0`'.

The simulation starts with generating all breakdown events and return from repair events for all components and placing them in the min-heap. This is done by the procedure ***generate_ttb_and_ret_from_repair***`(current_time,j)`, which takes as parameters the current time of the discrete-event simulator and the index 'j' of the component. The random time to a breakdown for the component with index 'j' is generated by a function which takes as parameter the breakdown frequency of the component with index 'j' and generates a random time to a breakdown by using the *inverse transformation method* [26] for sampling from the negative exponential distribution (1). This method consists of generating a random number x_i uniformly distributed in the interval (0,1) [27,28] and determining the corresponding random time to a breakdown t_i through the inverse function of the distribution of the time to a breakdown:

$$t_i = -\frac{1}{\lambda}\ln[1-x_i] \tag{5}$$

where λ is the breakdown/failure frequency of the corresponding component.

The actual time of the component breakdown is obtained by adding the current time of the simulator to the random time to a breakdown. Next, the procedure ***generate_ttb_and_ret_from_repair***`(current_time,j)` generates a random time to repair. A random time to repair is generated by taking the mean time to repair MTTR of the failed component and using the inverse transformation method to sample from the negative exponential distribution (2). Again, a uniform random number x_i is generated [27,28] in the interval (0,1) and the random time to repair t_i is obtained from

$$t_i = -MTTR \times \ln[1-x_i] \tag{6}$$

The return time from repair is obtained by adding the current time of the simulator, the random time to a breakdown and the random time to repair.

Checks are also performed whether the breakdown time and the time of return from repair are smaller than the length of the operation interval '`op_int`'. If this is the case, event records are made with the corresponding time stamps and are subsequently inserted in the min-heap. The min-heap data structure is contained in the array a[] whose elements are event records.

For the sake of simplicity, the implementation details related to inserting an event in the min-heap have been omitted here.

In the simulation loop, the nested while-loop goes through the events from the min-heap by always taking the event from the top of the mean-heap (Fig.5) with the statement '`ev=a[1]`'. This is the event with the smallest time stamp, therefore this is the event which will occur next. The current time of the simulator is updated with the earliest time by the statement '`current_time = ev.time_of_event`'.

Next, for the extracted from the min-heap event, a check is performed whether there will be available repairers to cover the demand for repair. If the remaining repairers are smaller than the required number of repairers, a constraint on the repair resources has occurred, the counter 'count' is inclemented and the while-loop is exited immediately by the statement '**break**'. This is done by the fragment:

```
if (rem_resources < ev.req_rep)
  {
    count=count+1;
    break;
   }
```

If the remaining repairers are sufficient to cover the demand for repair from the current breakdown event, the fragment

```
else {
   rem_resources = rem_resources - ev.req_rep;
   remove_event();
     }
```

allocates repairers to the breakdown event and decreases the amount of available repairers by the amount '`ev.req_rep`' required by the breakdown event. Next, the breakdown event is removed from the min-heap by calling the procedure '***remove_event***()'. Implementation details have been omitted here, for the sake of simplicity. New events with random breakdown time and random return from repair time are generated only when the current event is of type 'return from repair'. This is done in the fragment

```
else {
  rem_resources = rem_resources + ev.req_rep;
  remove_event();
  generate_ttb_and_ret_from_repair(current_time,
                                   ev.comp_indx);
    }
```

In this fragment, a release of engaged repairers is made first by the statement

```
rem_resources = rem_resources + ev.req_rep;
```

which is followed by removing the event from the min-heap by calling the procedure ***remove_event()***.

This is followed by calling the procedure ***generate_ttb_and_ret_from_repair***(current_time, ev.id), which generates new time-to-a-breakdown event and return-from-repair event for the corresponding component, with index 'ev.id'.

These steps are repeated while there are still events in the min-heap or the current time of the simulator is smaller than the length of the operational interval 'op_int'. The probability of constraint on the repair resources is obtained in the statement

```
prob_of_constr = count/num_simul_trials;
```

by dividing the number of times constraint on the repair resources has been registered (the content of the counter 'count') to the total number of simulation trials num_simul_trials.

Next, the fragment of the repeated bisection follows, from which the discrete-event simulator is called a number of times until the optimum is reached.

Finally, the algorithm for determining the optimal size of the system that can be serviced by a given number of repairers (so that the probability of constraint on the repair resources remains below a specified level) has also been implemented. This algorithm is very similar to the one described earlier and will not be presented here.

III. A Simulation Study and Results Related to a Large Computer System Experiencing Random Component Failures and Random Repair Times

In the simulation study, the random time to a breakdown was modelled by the negative exponential distribution (1) while the random repair time was modelled by equation (2). The data related to the number densities of the breakdowns related to the sparate devices and the mean time to repair have been taken from published studies ([11,13,19,20,21].

By using the described algorithm, the probability of constraint on the repair resources for a large computer system has been studied. The computer system included 11 different types of components: Workstations, WLAN transmitters, Printers, Servers; VoIP Servers, Routers, Switches, Fiber optical cables, Ethernet cables, Console cables and Monitoring computers.

Each of these devices is characterised its own breakdown/failure rate and mean time to repair. A single repairer was assumed for each of the failed devices.

The overall number of devices which experience breakdowns and require a repairer was 632. The simulation was run on a computer with 3 GHz Quad Core CPU. For 3 available repairers, a probability of constraint on the repair resources equal to 0.46 was calculated within 3.73 seconds, based on 10000 simulation trials. This demonstrates the high computational efficiency of the developed simulator based on a min-heap. The convergence was also very good because increasing the number of simulations to 100000 resulted in a very similar value (0.45) for the probability of constraint on the repair resources.

The probability of constraint on the repair resources as a function of the number of available repairers is given in Fig.6. As can be seen from the plot in Fig.6, there is a critical number of repairers for which the probability of constraint on the repair resources decreases sharply. For the considered computer system, this critical number is 4. For 3 repairers, the probability of constraint on the repair resources is 46%. Adding one more repairer however, causes the probability to drop sharply to 4.4%. The subsequent increase of the number of repairers causes only an insignificant decrease of this probability. The conclusion is that keeping more repairers than the critical number increases the salary costs without bringing a substantial decrease in the probability of constraint on the repair resources and an increase in the quality of service. Conversely, a number of repairers smaller than the critical number is associated with a large probability of constraint on the repair resources and a reduced quality of service.

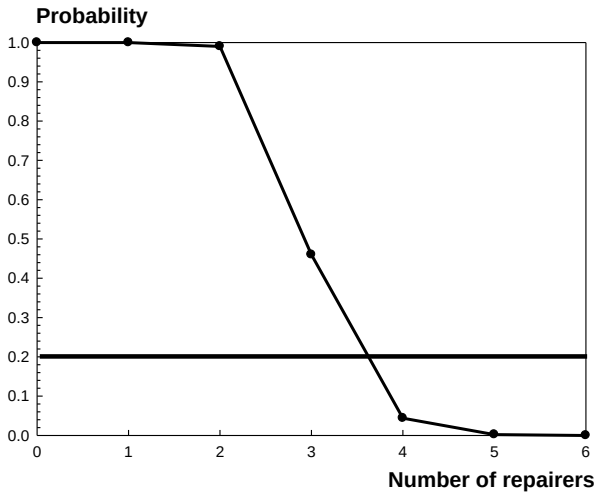


Fig. 6. Variation of the probability of constraint on the repair resources with the number of repairers

Determining the optimal number of repairers was conducted by using the same set of input data. The specified level of the probability of constraint on the repair resources was 20% ($\alpha = 0.2$).

The optimal solution was bounded between the values 0 and 30. Indeed, a number of repaires equal to zero yields 100% for the probability of constraint on the repair resources. A number of repairers equal to 30 yielded 0% probability of constraint on the repair resources. Consequently, the optimal solution must be located within the interval (0,30). Running the repeated bisection algorithm yield an optimal number of repairers equal to 4. The corresponding probability of running out of repair resources for this number of repairers is 0.044 (4.4%). The results from the previous simulation served as a validation test for the optimisation procedure. As it can be verified from Fig.6, four repairers is indeed the optimal number of repairers because 4 is the smallest number of repairers which yields a probability of constraint on the repair resources smaller than 20%. Three repairers yield probability equal to 0.46 - significantly larger than the specified maximum tolerable probability of 20%.

To determine the optimal size of the computer system which can be serviced by a given number of repairers, a weighted averaged breakdown frequency of 0.358 year^{-1} and a weighted average time to repair equal to 0.28 days has been used, taken from published breakdown data related to computer networks. Two levels of the maximum tolerable probability of constraint on the repair resources have been specified: $\alpha = 20\%$ and $\alpha = 2\%$. The optimal size of the network serviced by a different number of repaires, at the specified probability of constraint on the repair resources, is shown in Fig.7.

A serviced system size significantly smaller than the optimal size means that the available repairers are not used efficiently. A serviced system size significantly larger than the optimal size means that there is an increased probability of constraint on the repair resourecs which entails a low quality of service.

The developed discrete-event simulator also permits investigating the constraint on the repair resources related to specific components.

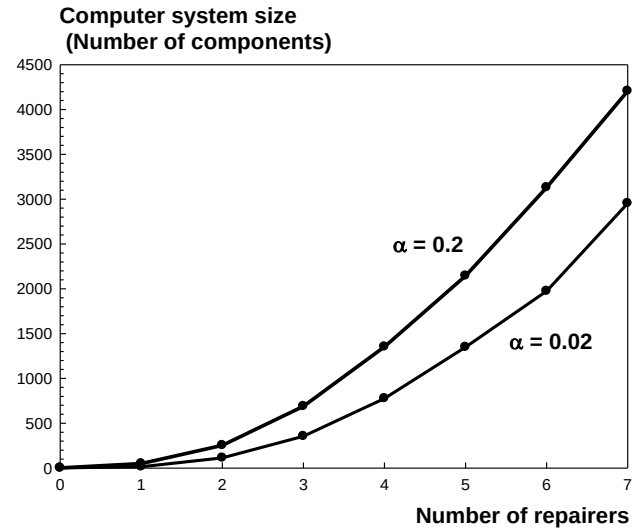


Fig. 7. Optimal serviceable size of the computer system for different number of available repairers

The failures of top of rack switches require the intervention of an operator. The failure frequency of top of rack switches according to [11] is $\lambda = 0.038$ year^{-1}. The cumulative empirical distribution of the time to repair is given in Fig.8.

The simulation based on 300 switches and 10000 simulation histories revealed a probability of constraint on the repair resources equal to 0.11. The time for computing this result on a computer with 3 GHz Quad Core CPU was 1.18 seconds. The empirical time to repair distribution in Fig.8 was sampled by combining the inverse transformation method and linear interpolation (the arrows in Fig.8).

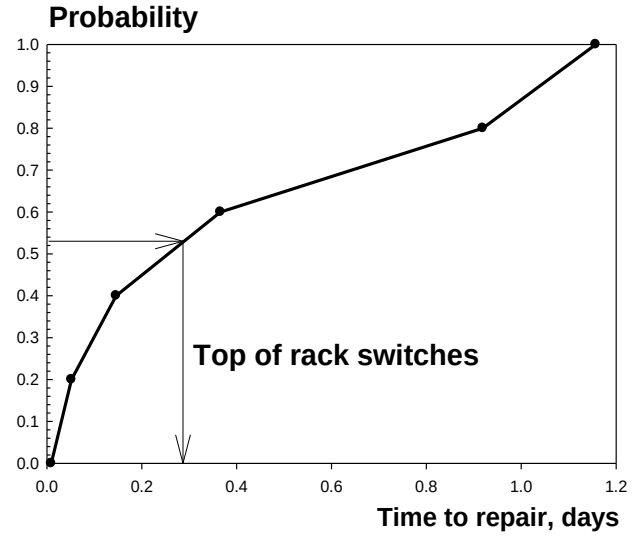


Fig. 8. Time to repair distribution of top of rack switches in computer systems (according to Gill et al.[11])

IV. Investigating the Influence of Critical Parameters on the Probability of Constraint on the Repair Resources

A. Investigating the influence of the failure frequency and mean repair time on the probability of constraint on the repair resources

The influence of the breakdown frequency on the probability of constraint on the repair resources has been tracked by keeping all the parameters constant except the breakdown frequency λ of the components, which was varied in the interval (0 -:- 4.0 year^{-1}), Fig.9. Ten components have been used, for each of which the same breakdown frequency λ has been assumed and one repairer was required for each component failure. For the purposes of the parametric study, the repair times have been assumed to follow a Gaussian distribution. A common mean time to repair MTTR=1.5 days has been assumed for each component, with a standard deviation σ equal to 0.15 days.

The operational interval was set to be 1 year. Two distinct cases have been investigated: (i) A single available repairer and (ii) two available repairers. The results have been summarised in Fig.9.

The results in Fig.9 reveal *a unexpectedly large probability of constraint on the repair resources* for a single available repairer. For the system including 10 components, each of which experiences on average 2 breakdowns a year ($\lambda = 2$, year^{-1}), with 1.5 days duration for repair, the probability of constraint on the repair resources during one year of operation is 73%! If the components experience $\lambda = 3$ instead of 2 expected number of failures per year, the probability of constraint on the repair resources is already 94%! These unexpected results show how easy it is to underestimate the probability of constraint on the repair resources. The result from such a underestimation are poor management decisions related to the number of people necessary to maintain a large system.

From the graphs, it can also be seen that increasing the number of repairers decreases significantly the probability of constraint on the repair resources. The significant decrease of the probability of constraint on the repair resources with the inclusion of a second repairer can be explained by the following.

If a single repairer is present, a constraint on the repair resources occurs whenever the repair times of two failed components overlap. The overlap means that a repairer is required at two different places (for two different failed components). The result is a constraint on the repair resources and a delay.

If two repairers are available, even if in the presence of overlapping repair times from two failed components, the repairers will work in parallel and no constraint on repair resources will occur. A constraint on the repair resources with two available repairers occurs only if three repairs simultaneously overlap (have a common point in time), the probability of which is relatively small. Providing an extra repairer brings a dramatic decrease in the probability of constraint on the repair resources.

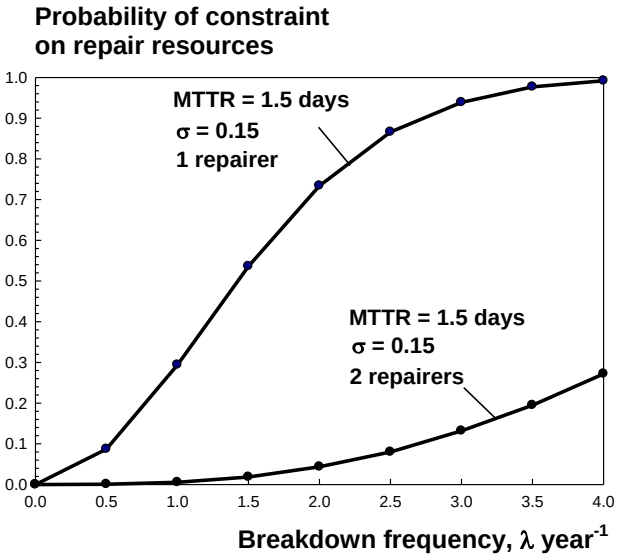


Fig. 9. Dependences of the probability of constraint on repair resources on the breakdown frequency of the components for different number of repairers

Next, the influence of the mean repair time on the probability of constraint on the repair resources was investigated by keeping all the parameters constant except the mean repair time. To separate the influence of the mean repair time from the influence of the standard deviation of the repair time, Gaussian distribution has been assumed for the distribution of the repair times. Note that for a negative exponential distribution of the times to repair, the mean and the standard deviation are both equal to MTTR [29], and their influences on the probability of constraint on the repair resources cannot be separated.

The mean repair time was varied in the interval (1-:-20 days; Fig.10). Ten components have been used, for each of which two different breakdown frequencies were assumed $\lambda = 1.0$ year^{-1} and $\lambda = 0.5$ year^{-1} and one repairer was required for each component failure. The number of available repairers was one; the standard deviation of all repair times was assumed to be constant: $\sigma = 0.15$ days. The results are presented in Fig.10.

With increasing the mean time to repair, the probability of constraint on the repair resources monotonically increases. Large times to repair yield a relatively small increase in the probability of constraint on the repair resources. No matter how large the mean time to repair is, there is always a non-zero probability that, within one year of operation, there will be no component breakdowns in the system. As a result, the probability of constraint on the repair resources tends asymptotically to unity. At a constant mean time to repair, the breakdown frequency has a strong effect on the probability of constraint on the repair resources.

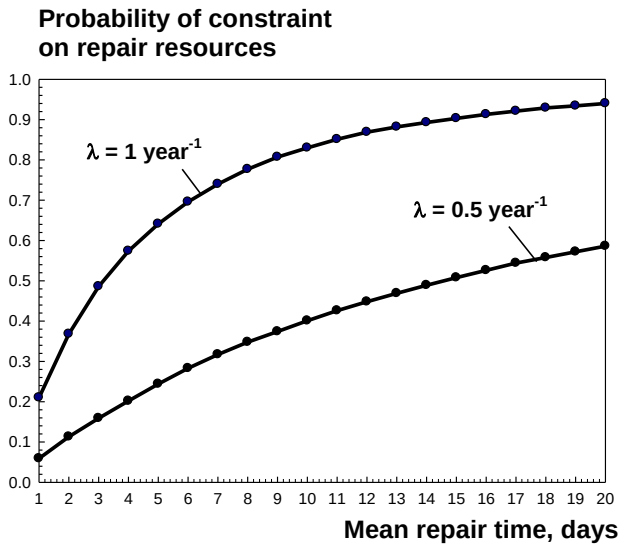


Fig. 10. Dependence of the probability of constraint on repair resources on the mean repair time

By using the discrete-event simulator, the influence of the standard deviation of the time for repair, on the probability of constraint on the repair resources, has been investigated. A system including ten identical devices, each characterised by a breakdown frequency $\lambda = 0.5$ year^{-1}, has also been simulated. Again, to separate the influence of the standard deviation of the repair time from the influence of the mean repair time, the times to repair of the devices have been assumed to follow a Gaussian distribution. The mean repair time was MTTR=60 days and the standard deviation of the repair time was varied from 0 to 10 days. The simulation experiments were repeated for one, two and three available repairers. The results indicated that the probability of constraint on the repair resources is insensitive to the variation of the standard variation of the repair time. These results showed that the probability of constraint on the repair resources cannot be reduced by reducing the variances of the repair times. The only way to reduce the probability of constraint on the repair resources is to reduce the mean repair times.

An important extension of this study is the analytical treatment of the problem related to estimating the probability of constraint on the repair resources. In addition, the developed methods could be applied with success to optimise the maintenance of production systems, transportation networks, water distribution systems and electrical distribution networks.

V. CONCLUSIONS

1) An optimisation algorithm has been proposed for the first time, for determining the optimal number of repairers maintaining a computer system of given size. The optimal solution guarantees that the probability of constraint on the repair resources will not exceed a maximum tolerable level.

2) An optimisation algorithm has been proposed for the first time, for determining the optimal size of a computer system which can be serviced by a given number of repairers. The optimal solution guarantees that the probability of constraint on the repair resources does not exceed a maximum tolerable level.

3) Both optimization algorithms are based on a repeated bisection. Unlike heuristic optimisation algorithms, the proposed algorithms are exact and always guarantee optimal solutions.

4) A very efficient discrete-event simulator has been created for the first time, for determining the constraint on the repair resources in large computer systems. The discrete-event simulator handles very large systems including thousands of components and is characterised by a very high computational speed.

5) The implementation of the discrete-event simulator is based on a min-heap data structure which guarantees that each operation involving inserting and deleting an event has a logarithmic running time. This running time associated with the insertion and removal of events is at the heart of the high computational speed of the developed discrete-event simulator.

6) The simulation results indicated the existence of a critical number of repairers at which the probability of constraint on the repair resources decreases sharply.

7) The parametric studies revealed an unexpectedly high probability of constraint on the repair resources for a single available repairer. This unexpected result shows how easy it is to underestimate the probability of constraint on the repair resources, which leads to poor management decisions.

8) The parametric studies indicated that providing more than a single repairer decreases dramatically the probability of a constraint on the repair resources.

9) Breakdown data and maintenance data related to computer systems should be component-specific. This provides the opportunity to use discrete-event simulators for optimizing the maintenance of computer systems.

10)An anonymized breakdown data base, similar to databases already existing for military electronic equipment, is a solid base for assessing and designing new computer systems and making correct management decisions.

11)This study could be extended by estimating the probability of constraint on the repair resources analytically. The developed methods could also be applied for optimising the maintenance of production systems, transportation networks, water distribution systems and electrical distribution networks.

REFERENCES

[1] J. Banks, J.S. Carson, B.L. Nelson, D.M. Nicol, Discrete-event simulation 4th ed., Prentice Hall, 2005.

[2] H. Al-Bahadili, Simulation in Computer Network Design and Modelling: Use and Analysis, Petra University, Jordan, 2012.

[3] M. Zukeman, D.Neame, R.Addie, Internet Traffic modelling and future technology implications, Proceedings of the 2003 InfoCom Conference, San Francisco, CA, 2003.

[4] P.Barford and M.Crovella, An architecture for a WWW workload generator, *Proceedings of the 1998 SIGMETRICS Conference*, Madison, WI, pp.151-160, 1998.

[5] N.I.Sarkar N.I., S.A. Halim, "A Review of Simulation of Telecommunication Networks: Simulators, Classification, Comparison, Methodologies, and Recommendations", Journal of Selected Areas in Telecommunications (JSAT), March Edition, 2011, pp.10-17.

[6] K.S.Trivedi, Probability and statistics with reliability, queuing and computer science applications, 2nd ed., Wiley, 2002.

[7] L.C. Wolstenholme, Reliability modelling, a statistical approach, Chapman & Hall, 1999.

[8] N.A.Weiss, *A course in probability*, Pearson Education, Inc., 2006.

[9] C.M.Carter, A.W. Malerich, The Exponential Repair Assumption: Practical Impacts, Proceedings of the Reliability and Maintainability Symposium, 2007. RAMS '07. Annual, Orlando, FL (2007).

[10] C.E.Ebeling, *An introduction to Reliability and Maintainability Engineering*, McGraw-Hill, (1997).

[11] P.Gill, N.Jain, N.Nagappan, "Understanding Network Failures in Data Centers: Measurement", Analysis, and Implications, SIGCOMM'11, August 15-19, 2011.

[12] D. Tang, R.K. Iyer, and S.S. Subramani., "Failure analysis and modelling of a VAX cluster system". In *FTCS*, 1990.

[13] R.K. Sahoo, A. Sivasubramaniam, M. S. Squillante, and Y. Zhang. "Failure data analysis of a large-scale heterogeneous server environment", In *Proc. of DSN'04*, 2004.

[14] J.S. Plank and W. R. Elwasif. Experimental assessment of workstation failures and their impact on checkpointing systems. In *FTCS'98*, 1998.

[15] D. Nurmi,, J. Brevik, and R. Wolski. Modeling machine availability in enterprise and wide-area distributed computing environments. In *Euro-Par'05*, 2005.

[16] B. Murphy and T. Gent. Measuring system and software reliability using an automated data collection process. *Quality and Reliability Engineering International*, 11(5), 1995.

[17] M. Kalyanakrishnam, Z. Kalbarczyk, and R. Iyer. "Failure data analysis of a LAN of Windows NT based computers", In *SRDS-18*, 1999.

[18] MIL-HDBK-217F, *Reliability prediction of electronic equipment*, US Department of Defence, Washington, DC, (1991).

[19] B. Schroeder., G.A.Gibson, "A large-scale study of failures in high-performance computing systems", Proceedings of the International Conference on Dependable Systems and Networks (DSN 2006), Philadelphia, June 25-28, 2006.

[20] B. Schroeder, G.A.Gibson, "The computer failure data repository (CFDR)", Workshop on Reliability Analysis of System Failure Data (RAF'07) MSR Cambridge, UK, March 2007.

[21] C. Labovitz C. and A. Ahuja. Experimental study of internet stability and wide-area backbone failures. In *The Twenty-Ninth Annual International Symposium on Fault-Tolerant Computing*, 1999.

[22] Failure data, 2006 (related to computer networks) http://www.pdl.cmu.edu/FailureData/ and http://www.lanl.gov/projects/computerscience/data/, 2006.

[23] T.Fujii and T.Dohi, Statistical analysis of a web server system, 2009 International Conference on Availability, Reliability and Security, pp.554-559, 2009.

[24] L.Leemis, L. M., Park, S. K. Discrete-event simulation: A first course. Upper Saddle River, N.J.: Pearson Prentice Hall (2006).

[25] T.H.Cormen, T.C.E.Leiserson, R.L.Rivest, and C.Stein, *Introduction to Algorithms*, 2nd ed., MIT Press and McGraw-Hill, (2001).

[26] S. Ross, *Simulation 2nd edition*, Harcourt academic press, 1997.

[27] L'Ecuyer, Efficient and portable random number generators, *Communications of the ACM*, vol.31, pp.742-749, 1988.

[28] Park S. and K.Miller, *Random number generators*, Commun. ACM, vol.31 (10), (1988), pp. 1192-1201.

[29] S.Ross, Introduction to probability models, 7th ed., Harcourt Academic press, 2000.

Supporting Self-Organization with Logical-Clustering Towards Autonomic Management of Internet-of-Things

Hasibur Rahman*, Theo Kanter, Rahim Rahmani
Department of Computer and Systems Sciences (DSV)
Stockholm University
Nod Building, SE-164 55 Kista, Sweden

***Abstract*—One of the challenges for autonomic management in Future Internet is to bring about self-organization in a rapidly changing environment and enable participating nodes to be aware and respond to changes. The massive number of participating nodes in Internet-of-Things calls for a new approach in regard of autonomic management with dynamic self-organization and enabling awareness to context information changes in the nodes themselves. To this end, we present new algorithms to enable self-organization with logical-clustering, the goal of which is to ensure that logical-clustering evolves correctly in the dynamic environment. The focus of these algorithms is to structure logical-clustering topology in an organized way with minimal intervention from outside sources. The correctness of the proposed algorithm is demonstrated on a scalable IoT platform, MediaSense. Our algorithms sanction 10 nodes to organize themselves per second and afford high accuracy of nodes discovery. Finally, we outline future research challenges towards autonomic management of IoT.**

***Keywords*—autonomic management; Future Internet; Internet-of-Things; self-organization; logical-clustering; MediaSense**

I. INTRODUCTION

Research towards Future Internet mandates exploring new challenges. Autonomic management has been around for over a decade since IBM coined this concept [1]. One of the requirements of autonomic management in Future Internet is to bring about self-organization [2, 3]. Moreover, we are about to see a paradigm shift from Internet-of-Things (IoT) to Internet-of-Everything (IoE) [4, 5]. The goal of which is to integrate people, process, data (context information) and things [5] in the Connected Society. Autonomic management was not part of early IoT, however, recently there is a shift in research activities which aims to tie these two [4]. This corresponds to massive participation of nodes in IoT, for example, 212 billion things are expected to be connected to IoT by 2020 [5]. The number of connected devices may even upsurge to 500 billion [6]. This massive immersion within IoT networking mandates to comprehend dynamism [7, 8]. One of the challenges in IoT would be to adapt to fast varying environment and be aware of any changes. The key to unravel these challenges is to organize a system such that it can respond to changing environment and stabilizes the system in situations uncalled for. For example, network connectivity, bandwidth, insertion and deletion of information, joining and leaving of a node/device, etc. are some of the changes that are expected in IoT [5, 8]. Any distributed system requires countering the changes in an organized way; however, most of these changes are not predictable. Therefore, it is imperative that the system organizes itself in such cases. This infers that a system (part of IoT) is desirable to be self-organized leaving outside sources mostly out of the loop.

The self-organization phenomenon exists in wide range of disciplines extending from physics to biology [9, 10]. It has also attracted attention from computer science and is now a very active research area [10]. Some case studies for self-organization in computer science have been presented in [9] which are inspired from the nature. This reflects the vision of autonomic computing that was coined by IBM [1]. They envisioned automatic computing as a grand challenge and outlined four aspects to be the core of automatic computing such as *self-configuration, self-optimization, self-healing, and self-protection* [1]. The requirement of these *self-** capabilities in massively distributed systems have been further discussed in [11]. The definition of self-organization varies in different disciplines befitting the respective goals and criteria. In general, self-organization can be considered as a system which organizes itself automatically i.e. without any intervention from outside sources [9, 12]. However, keeping outside source completely out of the loop is still a research challenge. Even the vision of autonomic computing states: "*a system should organize itself according to high-level objectives, and will collect and aggregate information to support decisions by human administrators*" [1]. This has further been outlined as "*Put simply, the autonomic paradigm seeks to reduce the requirement for human intervention in the management process through the use of one or more control loops that continuously reconfigure the system to keep its behavior within desired bounds*" [13].

Franco and Omer in "*IEEE Transaction special issue: Self-Organization in Distributed Systems Engineering*" have further stressed that self-organizing systems shall leave human mostly out of the loop [14]. They have wisely used the word "*mostly*", because although it is desirable but currently it is impractical to leave outside sources (e.g. human as an administrator- these will be used interchangeably in rest of the paper) completely out of the loop. This implies that a system should execute its tasks even when there is minimal or no support at all from outside source. But the system should be able to interact with outside source and run a periodic algorithm to rectify errors

and make system aware and learned about faults. Thus, the system will be able to evolve itself next time it encounters similar error and, thereby, reducing the outside intervention as much as possible. Therefore, self-organization can be defined as a system that "*evolves correctly, adapts to dynamic situations, stabilizes itself in unpredicted situations, and pre-protects itself against probable attacks*".

In light of above, this paper aims to design and develop new algorithms to empower self-organization. The algorithms will be specifically targeted at logical-clustering approach. Logical-clustering efficiently filters out similar context information from distributed sources [15]. MediaSense, a context sharing Internet-of-Things (IoT) platform [16], has been employed to disseminate the clustering identity (context-ID) as a PubSub model for logical-clustering [17]. Results showed that MediaSense is viable for scalable context-ID distribution. It has been further demonstrated in [8] that MediaSense can counter the fast varying environment i.e. dynamism efficiently as well. One of the focuses of this paper is to address "*how can logical-clustering organize itself and evolves according to the requirement and manages in dynamic unpredictable circumstances?".* This paper is particularly interested in automatic and periodic insertion and deletion of context-IDs; self-configuration (automatic seamless integration of participating nodes) and self- healing of sinks (nodes and sinks are used interchangeably throughout this paper); self-optimization of both context-IDs and sinks etc. The correctness of newly designed and developed algorithms will be proven on MediaSense. As it has been already proven that MediaSense can easily counter scalable distribution of context information and it can support dynamism, therefore, it is worthy that we develop self-organization algorithms using MediaSense. This will also contribute in structuring logical-clustering topology in an organized way.

The remainder of the paper is structured as follows: section II presents the motivation, section III revisits the state of the art autonomic computing, section IV demonstrates our approach while section V demonstrates the evaluation of the work, section VI outlines future research challenges, and finally section VI concludes the paper.

II. MOTIVATION

The idea of autonomic computing has been around over a decade now. Since then, there have been several proposals for self-organizing systems. Our goals in this paper are: to support self-organization with the logical-clustering and to propose algorithms which should contribute towards an autonomic IoT management architecture. The aim is to bring about self-organization to the logical-clustering approach. The overall objectives are to ensure that logical-clustering evolves correctly in dynamic and unpredictable situations, and structures itself in an organized manner. As logical-clustering implies that context-ID (identification of cluster) is created as soon as a cluster is formed and depending on the requirement (policy) context-IDs should be deleted after a specific time. The system needs to be aware of this i.e. *self-optimized*. Each sink in logical-clustering needs to fetch context-IDs from other sink(s). The system should be configured in such a way that sink(s) can fetch CI from other sinks automatically and periodically. This also implies that sink should discover other available sinks and advertises itself so that seamless integration to the system is ensured. This mandates establishing *self-configuration* as outlined by autonomic computing vision. Sink in logical-clustering topology is considered to be fixed. And fixed sinks are considered not removable; however, sinks can be down for example due to no Internet connectivity, power failure, etc. Hence, it is imperative the system should re-configure whenever it is up again. Re-configuration also includes retrieving old data and synchronizing immediately with other sinks automatically. This is branded as *self-healing*. Therefore, our particular focus is limited to design and develop algorithms for *self-configuration, self-optimization and self-healing*. The algorithms will be evaluated on a proven, scalable, and adaptable IoT platform MediaSense.

III. AUTONOMIC MANAGEMENT

This section briefly introduces the state of the art autonomic computing and further reviews each of the fundamental aspects of self-organization system.

A. Revisting the state-of-the-art

A self-organized system deemed to be dynamic and each organized element constitutes overall system, thus, the resulting overall system becomes complex and its behavior becomes unpredictable [11]. According to many research papers, as mentioned earlier, a self-organized system should acquire its organized characteristics without intervention from outside sources [9, 12]; however, some other researchers mention that outside sources should be kept out of the loop as much as possible [13, 14]. The vision of autonomic computing would only be complete through acquired knowledge i.e. awareness from each organized system [1, 13]. Therefore, the overall system should evolve gradually and eventually keep outside sources out of the loop mostly- if not completely.

Self-organized system can have several self-* capabilities [11]. All these self-* capabilities can be summed into the four aspects of autonomic computing i.e. self-configuration, self-optimization, self-healing, and self-protection. Table I further illustrates this [1].

These aspects need to be managed by a manager which will enable interaction with other organized elements and/or outside source (e.g. human administrator). Manager will enable analyzing, planning and executing the high-level objectives (policies) set by outside sources and it will further allow an organized element to interact with another organized element inside the system. Fig. 1 shows how a manager can achieve this (the figure is redrawn from [1]).

TABLE I. SELF-* ASPECTS

Aspects	Capabilities
Self-Configuration	A new node joins; advertises itself and discovers other nodes Adjusts and integrates automatically and seamlessly according to the high-level policies
Self-Optimization	A node should be able to optimize the local operation parameters according to global objectives Learning and altering objectives adpated by others Should be able to adjust in case of policy conflict
Self-Healing	Re-configurations of the nodes in case of failures Healing for configuration and optimization
Self-Protection	A node should be able to protect itself from outside and undesirable attack

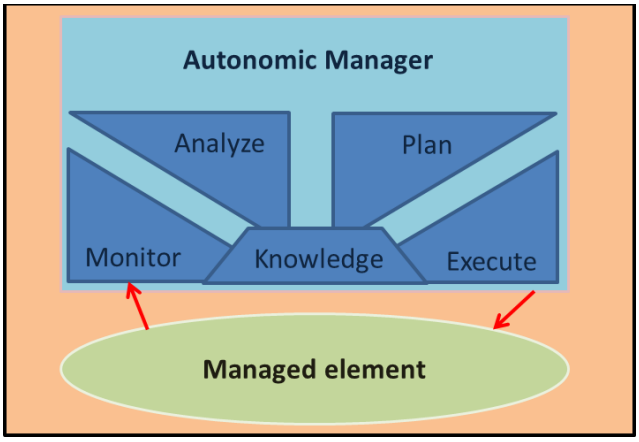


Fig. 1. Structure of an autonomic element [1]

An element joins the system and the autonomic manager, first of all, analyzes the element, then plans and finally executes based on the objectives required by the overall system. This behavior of joining and execution follows the trend of a control loop (the red arrows correspond to this). That means autonomic computing relates to a control loop which is executed based on the specified policies. Fig. 2 further demonstrates this. The policies (objectives) are responsible for implementing the self-* capabilities. At the beginning, these policies are integrated with the system; and as system evolves and encounters new problems, new polices are added. However, this requires learning i.e. awareness (through acquiring knowledge) from each organized element.

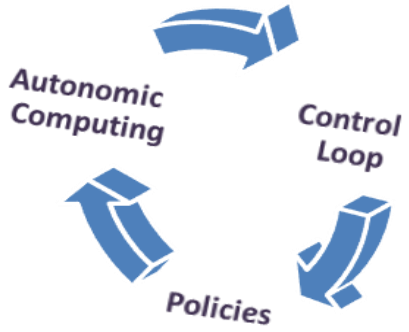


Fig. 2. Life cycle of autonomic computing

IV. OUR APPROACH

Our focus in this paper is not to redefine the self-* aspects outlined in previous section; rather we focus to design and develop algorithms that would implement these aspects. A limitation of this paper is that it will not explore the self-protection aspect. In this section, we will explain the proposed algorithms for other three self-* aspects targeted at logical-clustering. Moreover, the correctness of the algorithms will be demonstrated on a p2p based IoT platform- MediaSense.

MediaSense disseminates context information using Distributed Context eXchange Protocol (DCXP) where each entity is registered as Universal Context Identifier (UCI) and other entities can resolve the UCI [16]. It utilizes rendezvous host i.e. a bootstrapping node to initiate communication between entities. Any entity plans to use MediaSense must use the primitive functionalities defined in the MediaSense Platform. Our proposal is to utilize MediaSense Platform as controller i.e. autonomic manager and each of the designed three self-* aspects has been added as extended primitive functions for MediaSense. Fig. 3 shows how MediaSense Platform can be utilized as an autonomic manager. In the following sub-sections, we will describe how self-organization can be supported in logical-clustering by employing the autonomic computing concept. We have shown in [17], how MediaSense can be utilized for logical-clustering concept.

A. Self-Organization Support for Logical-Clustering

Logical-clustering consists of logical-sinks [15] and sinks are responsible for controlling (creation, insertion, deletion) the clusters. However, sinks require discovering, co-operating with other available sinks, and it should also maintain itself. In other words, it should be organized and organization should be done with minimum support from outside sources. Since logical-clustering involves real-time communication, it is imperative that it evolves correctly, automatically in real-time.

Fig. 3 shows how we want to employ autonomic computing concept i.e. self-organization in logical-clustering. As in any other system, a sink first needs to join which is the first operation in the control loop.

An autonomic manager (i.e. MediaSense Platform in our case) will analyze the policy associated with the join request. During this operation, MediaSense platform will implement the self-* algorithms. The sink will be adapted based on the outcome of the algorithms after which sink would be ready for execution. The platform should be aware all of these operations as indicated in fig. 3. Actions showed in Fig. 3 can be summed up as following:

- *A sink joins*
- *Platform analyzes the policies i.e. evaluates the self-* algorithms*
- *Platform further adapts the sink based on the policies' outcome*
- *Sink is ready for execution (after this stage sink is said to be an organized sink)*
- *Platform has the awareness of all these actions*

Fig. 4 shows how the concept can contribute towards autonomic management of IoT. An entity i.e. a thing in IoT will be connected to the scalable MediaSense Platform and will be managed by the self-* algorithms automatically.

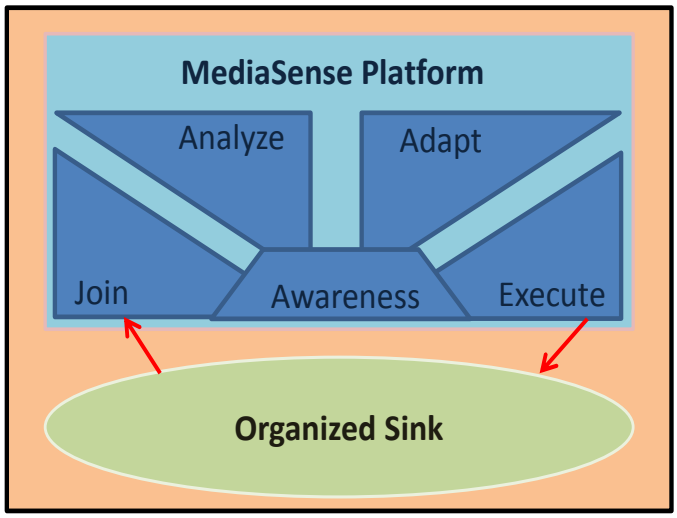


Fig. 3. Supporting self-organization with logical-clustering

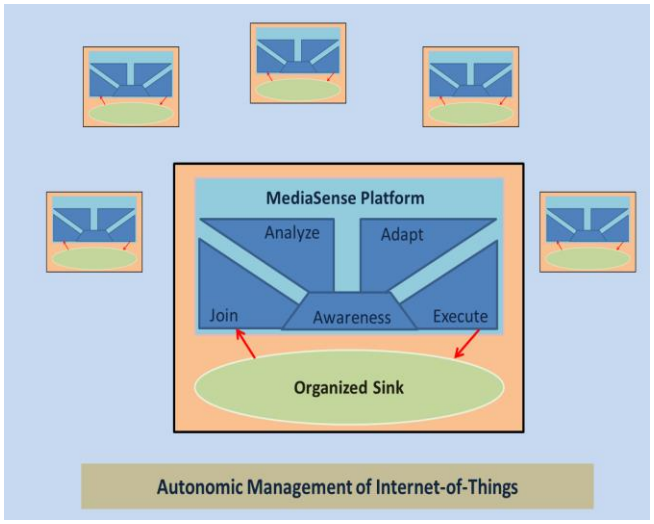


Fig. 4. Autonomic management of IoT

B. Self-Configuration

Self-configuration implies that whenever a new element/node joins a system; it should be able to advertise itself and discover other available nodes or let other discover inside the system. The goal of this self-* aspect is to ensure automatic and seamless integration to the joined system.

To achieve the aforementioned goal, we have employed publish/subscribe (PubSub) model that was shown in [17]. MediaSense utilizes a bootstrapping node, as the case with other distributed p2p systems, which needs to be initiated before any communication can take place. Our idea is to define a global publisher e.g. global_uci in MediaSense and each node whenever tries to join the MediaSense Platform automatically subscribes to the publisher without knowing which node holds the global_uci. This newly joined node is added to the global_uci and existing nodes are automatically discovered after a certain time. Self-configuration involves several steps. The steps are listed below:

- *Sink join*
- *Configure global_uci (at beginning on MediaSense bootstrap node)*
- *Sink configuration*
- *Check for self-healing i.e. reconfiguration*
- *Discover other sinks*

Each of the algorithms is depicted in the following figures.

```
Algorithm registering global_uci

begin
  initiate bootstrap_MediaSense
  //the followings are added as extension
to current MediaSense
  if global_uci exists
    renew global_uci
  elseif
    register global_uci (based on PubSub's
publisher algorithm)
  endif
  //the above is extension
  run bootstrap_MediaSense
  start P2P Communication on the bootstrap
IP address and bootport
end
```

Fig. 5. Algorithm for configuring a global_uci as publisher

Fig. 5 depicts the algorithm for configuring global_uci which is exploited as a publisher. Whenever MediaSense is bootstrapped, global_uci is either merged or renewed based on the requirement (different situation might require different policy). As for this approach, global_uci is renewed meaning that a fresh copy of global_uci is guaranteed. Global_uci stores all the available sinks as published items and each node - that joins a MediaSense platform and invokes the MediaSense Platform's selfConfig primitive function (fig. 7)- is automatically subscribes to the global_uci.

```
Algorithm sink_join

begin
  create an instance of MediaSensePlatform
  initialize platform with network settings
(Bootstrap IP address, bootport, local port)
  while MediaSense is bootstrapped
    declare UCI (i.e. identity of the node)
    invoke MediaSensePlatform's joinUCI
     if MediaSensePlatform's selfHealing is
true
      if the UCI is not listed on global_uci
       publish on global_uci by invoking
MediaSensePlatform's config method
       return the UCI's current configuration
status
      endif
     elseif
      register the UCI on MediaSense platform
      publish on global_uci by invoking
MediaSensePlatform's config function
     endif
    invoke MediaSensePlatform's
selfConfiguration
    synchronizes with the existing UCIs after
every T seconds
  endwhile
end
```

Fig. 6. Algorithm for sink joining

As mentioned earlier, each sink that joins the MediaSense Platform is identified as UCI. Therefore, every time a new sink joins, it fetches all the available UCIs i.e. sinks. Fig. 6 shows the algorithm for sink joining. In this procedure, first a sink must initiate the MedaSense Platform along with the network settings to use its functionalities.

An identity for this sink is declared which is then used to join the platform. When the joinUCI function is called- it first checks if UCI already exists (i.e. registered with MediaSense) with this particular sink- if that check returns true then sink is reconfigured with previous configuration (see fig. 8), if that check returns false meaning no identity duplication from this sink then the UCI gets registered on MediaSense platform and gets published on the global_uci which allows other sinks to discover node. The joined sink then invokes the selfConfig function of MediaSense Platform which enables automatic discovery of other available nodes on MediaSense every *T* seconds.

The interval is implementation dependent- by default set to 20 seconds. The interval is an open research issue for battery powered devices, rechargeable devices; but not an issue for AC powered sources. Discovery of sinks involves the idea that of MediaSense's subscription [17]. In this procedure, first the global_uci is resolved based on MediaSense's subscription algorithm. When the global_uci is resolved, each of the stored UCIs is read and current UCI is added to the global_uci. After adding the current UCI, global_uci is updated. This implies that global_uci is either merged or renewed based on the configuration requirement. Subscription is matched whenever this function- implementing this algorithm- is invoked.

```
Algorithm sink_discovery

begin
   resolve global_uci (based on PubSub's
subscription algorithm)
   read the subscribeable UCIs
   store the UCI to the global_uci
   update global_uci
   merge or renew depending on the
configuration
    subscription is synchronized whenever this
method is invoked
end
```

Fig. 7. Algorithm for sink discovery

```
I. Algorithm sink_duplication_check

begin
  resolves the UCI
  fetches associated information
  if the sink id is found with the fetched
information
   returns true
  elseif
   returns false
  endif
end


II. Algorithm sink_reconfigure

begin
  resolves the UCI
  fetches associated information
  if uci exists in the current MediaSense
   reconfigure the node and fetch previously
existing data
  elseif
   start uci registration
   while register
     invoke selfConfiguration
   endwhile
  endif
end
```

Fig. 8. Algorithm for sink reconfiguration

C. Self-Healing

Self-healing refers to the reconfiguration of a node in case of failure. This also corresponds to the healing for the other two self-* algorithms. The self-healing function that is called when a node joins checks if UCI is already exists. Moreover, this algorithm should ensure when a sink is down for some reason, it should be able to reconfigure with previous configuration whenever it is up again. Fig. 8 shows the algorithm. The first part of algorithm requires two parameters i.e. the UCI and sink-ID; and returns a boolean value. Sink ID is obtained by calling MediaSense Platform's getHostID function. This procedure begins by resolving the UCI and it fetches the associated information with this UCI. If the host ID is found in the fetched information then this function returns true otherwise a false value is returned. Second part of self-healing algorithm also requires resolving the UCI and its associated information are fetched. After that, if UCI exists on MediaSense then the algorithm reconfigures the sink with previously existing data. But if the UCI is nonexistent then UCI is registered on the Platform and selfConfiguration is invoked so that the UCI being registered can execute the self-configuration algorithm as discussed in previous sub-section.

```
Algorithm sink_optimization

I. Insert context-ID

begin
  resolves the UCI
  checks for new context-IDs
  if new context-IDs are found
    insert new context-IDs in the UCI and
adjusts seamlessly with existing context-IDs
    invoke Insert ContextID Policies
  endif
 end

II. Delete context-ID

begin
  resolves the UCI
  checks for context-IDs to be deleted
  if context-IDs need to deleted
    delete existing context-IDs in the UCI
and adjusts seamlessly with existing context-
IDs
    invoke Delete ContextID Policies
     if single context-ID with a TTL
      delete context-ID
     eleseif
       delete context-IDs
     endif
  endif
end
```

Fig. 9. Algorithm for sink optimization

D. Self-Optimization

Self-optimization implies that a node should optimize itself according to the policies set by manager (in our case MediaSense Platform). This should further ensure that each node performs to its best capability and efficiency. In this paper, we looked into optimizing context-IDs associated with logical-sink. The goal of this algorithm is to support autonomic dynamism in logical-clustering. This means automatic insertion, deletion of context-ID is made possible which enables awareness in the sinks. Each sink creates new context-ID in real-time; therefore, it is imperative that sink optimizes itself by inserting new context-IDs with an interval of *T* seconds. Each sink should also be able to delete context-IDs automatically whenever needed. Moreover, each sink should be aware if a particular context-ID is needed to be deleted after a specified time. Fig. 9 depicts the algorithm. This algorithms also has two main parts i.e. insertion and deletion of context-IDs. In the first part of this procedure, the algorithm first resolves the UCI and checks if there are any new context-IDs to be inserted.

When there are next context-IDs found for insertion, the class that implements the Insert ContextID Policies gets invoked. Depending on the requirement, different policies may be executed. In our case study i.e. logical-clustering, it automatically and periodically inserts the new context-IDs, integrates seamlessly with existing context-IDs and synchronizes with other sinks. As for deleting context-IDs, the procedure checks if a particular context-ID (with a timestamp for time to live (TTL) is attached) is to be deleted or it should delete context-IDs. The procedure executes these and optimizes the sink.

V. Performance Evaluation

This section presents the performance evaluation of supporting self-organization with logical-clustering. Self-* capabilities algorithms have been developed on the MediaSense platform. Each of the self-* aspects has been developed and included as separate package, this can be found under the new package called autonomic in current MediaSense platform. We start first with reporting the effect of incorporating self-organization on MediaSense Platform. Table II and II report this. We have evaluated the performance of joining node on two different networks with different Internet speeds. Measurements are shown in logarithmic scale and in milliseconds; the mean μ and standard deviation σ values are depicted in tables. Results suggest that on both networks, we get almost similar result. In terms of self-organization, we have divided the evaluation into two. In the first part, we do not consider time required for self-healing i.e. duplication check and reconfiguration; and only time required for self-configuration is considered, and the second part includes time required for self-healing (see IV-B). MediaSense incurs a delay if self-* algorithms are employed. This, however, is what is expected of self-* algorithms, reason being a node goes through the life cycle of autonomic computing (see fig. 2 & 3) before completing the joining operation i.e. becoming organized (a managed node). The algorithms demonstrated almost identical performance on both networks.

TABLE II. Node Joining (Network I)

	MediaSense (Current)	MediaSense (Self-Configuration)	MediaSense (Self-Organization)
μ	1.59	1.69	1.88
σ	0.0522	0.0459	0.0338

TABLE III. Node Joining (Network II)

	MediaSense (current)	MediaSense (Self-Configuration)	MediaSense (Self-Organization)
μ	1.61	1.69	1.91
σ	0.0593	0.0261	0.0260

First operation in the self-organization starts with node(s) joining the autonomic management platform. For this particular evaluation, we consider that nodes are not competing for MediaSense Platform's resources. This means that nodes are executed one after another. The issue of concurrent node(s) joining leads to load-balancing and scheduling issue which is not covered in this paper. Table VII elaborates the necessity for load-balancing and scheduling. Fig. 10 shows the performance of nodes joining the platform. X-axis represents the number of nodes and y-axis represents the processing time. Processing time is shown in logarithmic scale and in seconds here. If we analyze the figure and the processing time in normal scale, we find that each second 10 nodes can join the MediaSense Platform. This could be a starting point for exploring the load-balancing and scheduling issue. However, this result should not be confused with table II and II. Table II and II only reported the operation time require for a single sink joining operation, these did not consider the inside mechanism requires for fully organizing with other nodes (time was shown in logarithmic scale there too).

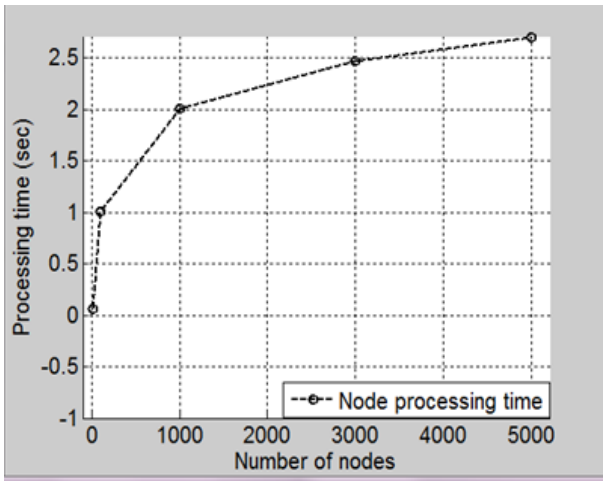


Fig. 10. Processing time for nodes joining

Next we evaluate the self-organization algorithms in terms of discovery of nodes and accuracy of discovery. Fig. 11 shows the discovery of nodes. This implies the time required for a node to discover i.e. synchronize with other nodes. This has been evaluated for both dynamic and stable scenarios. Dynamic means discovery of nodes measured while other nodes are joining simultaneously and stable implies that currently no more nodes are joining the system. This measurement was done when evaluating fig. 10. This algorithm is run after every 20 seconds. Table VI illustrates the algorithm depicted in fig. 7. The stable scenario corresponds to this. The results are different from dynamic scenario where each node goes through the cycle of self-organization and competes for resources; thus incurs delay. The result portrayed in table VI suggest that discovery of nodes while system is stable is very fast. This also corresponds to the MediaSense's PubSub model which also demonstrated fast and efficient result [17]. Hence, we do not discuss this in detail in the paper.

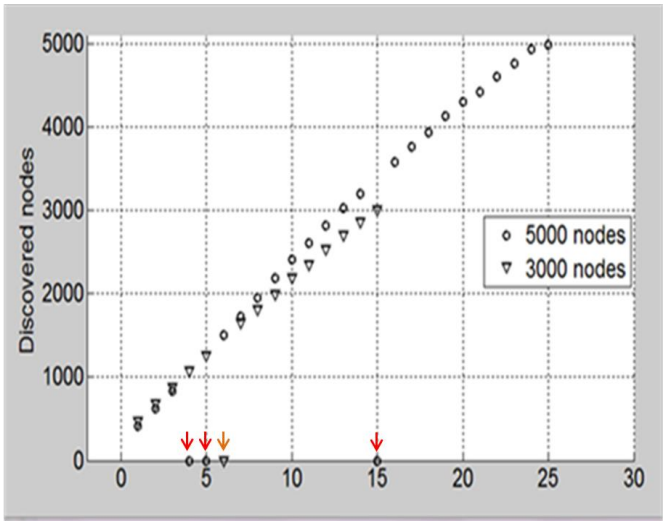


Fig. 11. Discovery of nodes

The above figure depicts nodes' discovery for 5000 and 3000 nodes joining the platform. The mean μ for discovery is 197, 184; and most frequent number (F) is 205 and 202 respectively for 5000 and 3000 nodes. These values have been calculated based on the differences of the vector elements excluding the first and last element that makes the discovery stable i.e. no more nodes are joining. This reflects that each second (in normal scale) 10 nodes join the platform. Since the interval period T was chosen to be 20 seconds, so the mean values are justified. This could further be seen from the fig. 11 that node discovery became stable (all available nodes were discovered) on the 15th (3000 nodes case) and 25th (5000 nodes case) attempt. However, 4990 and 2997 nodes out of 5000 and 3000 nodes could be discovered respectively. This gives a high discovery accuracy of over 99 %. Discovery accuracy can further be seen from table V. Also on the 4th, 5th and 15th attempt (5000 nodes case- red arrows) and on 6th attempt (3000 node case- orange arrow) of self-configuration, the algorithm failed to discover any nodes. This could be because currently the algorithm tries to renew (see fig. 7) the global_uci, this means the old global_uci is deleted and a fresh copy of global_uci is inserted on nodes. And, during the call and renew, some UCIs might not have been synchronized. But the system was able to fetch information immediately in next attempts meaning stability is ensured (further discussed below).

TABLE IV. NODE DISCOVERY (DYNAMIC)

	5000 nodes	**3000 nodes**
μ	197	184
σ	29.79	14.8613
F	205	202

TABLE V. DISCOVERY ACCURACY (SEQUENTIAL)

	5000 nodes	**3000 nodes**	**1000 nodes**
Nodes discovered	4990	2997	1000
Accuracy	99.8 %	99.9 %	100 %

Table V indicates that discovery accuracy is very high and near to 100 % for all three simulated cases, however, these measurements were done when nodes were joining sequentially. We did not consider concurrent nodes joining where nodes would compete to access MediaSense Platform resources. Table VII shows the discovery accuracy for 3000 and 2000 concurrent nodes joining. Arithmetic mean is 1699, 1195 and standard deviation is 110.4513, 97.4664; and the discovery accuracy drops to 56.63 % and 59.75 % respectively. This drop in discovery accuracy further highlights necessity for designing and developing load-balancing and scheduling algorithm.

TABLE VI. DISCOVERY DURATION (STABLE)

	5000 nodes	**3000 nodes**	**1000 nodes**
μ	54.67 ms	47 ms	44 ms
σ	9.76 ms	9.18 ms	9.57 ms

TABLE VII. DISCOVERY ACCURACY (CONCURRENT)

	μ	***σ***	***Accuracy***
3000 nodes	1699	110.4513	56.63 %
2000 nodes	1195	97.4664	59.75 %

Self-healing algorithm implies that each node should be able to heal itself and compensate for failure. The developed algorithms offer duplication check and compensate when a failed (e.g. was down or offline) node rejoins the system. This algorithm was evaluated with 2000 and 3000 already existing nodes trying to rejoin the system concurrently (this was evaluated while measuring for table VII) and each node was successfully checked for duplication. This means duplication check accuracy was 100 % for both cases. This was further confirmed with 1000 existing nodes trying to join one after another i.e. sequentially, this too offered 100 % success rate.

One of the goals of this paper was to make sure logical-clustering evolves correctly and automatically i.e. optimize itself.

Optimization still needs improvement (should be developed upon more policies being explored); as of now this algorithm (along with other two algorithms) allows us to structure the logical-clustering topology. A sink in logical-sink requires synchronizing with other sink(s). Thus far, this was done manually as shown in [8, 17]. Now it is possible to synchronize the logical-sink automatically and periodically. According to the experiments done, the algorithm has successfully achieved automatic dynamism (insertion, deletion) of context-IDs.

A self-organized system is considered complex and its outcome is often unpredictable [11]. Results demonstrated in this paper also reflect this. This can be easily perceived from the fig. 11, table IV, V and VI. Discovery nodes per attempt are not always same (see fig. 11); even the μ and σ do not exhibit similarity for both 3000 and 5000 nodes. However, by perceiving the standard deviation, the deviations are not too fluctuated and remain within reasonable limit.

Network stability and resilience are two of the main components of a system. Our self-organized algorithms were able to structure the logical-clustering topology and the topology was able to evolve correctly without any central point of failure. We have observed in fig. 11 that node failed to discover nodes in few cases, but it was able to stabilize itself immediately. Since global_uci holds the information about each existing sink and this information is accessible to each node subscribed to this global_uci. This makes the system resilient to failure i.e. no central point of failure. When a sink failed or left the network and attempted to rejoin the system, all previous information was fetched immediately. Fig. 11 further confirms this when after failing to discover in few cases, it was able to fetch old and new information simultaneously. Self-healing's 100 % success rate also reflects to such claim of stability and resilience.

VI. Future Research Challenges

The idea of logical-clustering was proposed in [15] and the research reported in this paper is a step forward to fulfilling the based vision. However, the vision of fully functional self-organized logical-clustering still requires some considerable research work. So far, we have not discussed the policies in organizing logical-clustering. In this paper, we have presented a template for achieving self-organization. A fully self-optimized system requires adapting new policies, and optimizing the system accordingly. Moreover, autonomic management of IoT would only be possible through exploring further policies. This mandates exploring the policies that would enable autonomic management of IoT and thereby IoE. Furthermore, integrating polices into the manager mandates a more flexible and concrete manager. This could be achieved by deploying Software-Defined Networking (SDN) concept. The intelligence of SDN would enable to counter the challenges of efficient traffic management, and data and services delivery.

Another fundamental aspect of self-organization i.e. self-protection also need to explored and implemented. Each node should be protected against possible attacks and from getting removed by other node(s). Protections of context-IDs also need to be ensured.

Our focus in this paper was limited to designing and developing a template for self-* aspects, and these self-* aspects need to be adapted based on awareness. This awareness should come from all the stages as depicted in fig. 3. Learning plays an integral part in building the awareness. The managers should be able to learn new policies and create awareness of the learned policies to other nodes. However, we have not addressed the issue of learning in this paper. An approach for learned-manager is still an open issue which can be implemented along with SDN and learned-management system.

Unique identification of each node is another important research issue needs to be addressed. However, uniquely identifying billions nodes is not something easy to implement. Moreover, these identifications should also be easy for humans to remember. For example, context-IDs in logical-clustering should be unique and human should be able to access these context-IDs easily. Therefore, it mandates to define a naming scheme which will ensure unique identification of node in IoT landscape.

Load-balancing and scheduling of autonomic manager is another feature that needs attention.

Heterogeneous interoperability of the system also remains a challenge. IoT heavily involves cross-platform communication, therefore, it is mandated that we look into cross-platform behavior of the system too.

VII. Conclusion

The contribution of this paper by and large lies with designing and developing the self-* aspect capabilities inspired by the autonomic computing concept. In particular, *self-configuration, self-optimization and self-healing* algorithms were designed and developed; and correctness of these algorithms was proven on a scalable and versatile IoT platform MediaSense. MediaSense Platform was employed as what is autonomic manager to autonomic computing. This new algorithms enable logical-clustering to organize automatically and periodically. Each sink in logical-sink now said to be organized and it evolves correctly.

The algorithms sanction 10 nodes to self-organize themselves in each second on the MediaSense Platform, and discovery accuracy is over 99 % when there is no competition for MediaSense Platform's resources. While nodes compete for MediaSense Platform's resources, discovery accuracy is around 60 %. Stable system allows discovering nodes very fast; and duplication check always succeeded. This enables logical-clustering topology to evolve correctly and structure itself. There is no central point of failure, even if bootstrap node fails, other node takes over and stabilizes the system.

Our next step is to explore the challenges mentioned in the future research challenges. SDN would, perhaps, enable us to see fully operational autonomic management of IoT. Autonomic management itself is a grand challenge and it will go through many transitions. IoT would also need to see-off many transitions and finally embrace an operational autonomic IoT- thereby IoE.

We hope that the algorithms depicted in this paper are step forward towards the autonomic management of IoT and could be used as template for further development.

ACKNOWLEDGMENT

The work is partially supported by funding from the European Union FP7 MobiS project. We would also like to thank our colleagues at Immersive Networking Research Group for their feedback that helped in finalizing paper.

REFERENCES

[1] J. 0. Kephart and D. M. Chess, "The vision of autonomic computing", IEEE Computer Society, 36(1):41-52, 2003.

[2] J. Strassner, S.S. Kim, J. Won-Ki Hong, "The Design of an Autonomic Communication Element to Manage Future Internet Services", Management Enabling the Future Internet for Changing Business and New Computing Services Lecture Notes in Computer Science Volume 5787, 2009, pp 122-132

[3] Siekkinen et al., "Beyond the Future Internet – Requirements of Autonomic Networking Architectures to Address Long Term Future Networking Challenges", IEEE computer society

[4] Wikipedia, "Internet of Things" http://en.wikipedia.org/wiki/Internet_of_Things [Last Accessed: 04-February-2015]

[5] Ruthbea Yesner Clarke, "Smart Cities and the Internet of Everything: The Foundation for Delivering Next-Generation Citizen Services", white paper sponsored by Cisco, October 2013

[6] Ericsson White Paper, "5G Radio Access: Research and Vision", 2014

[7] A. Zaslavsky, "Adaptibility and Interfaces: Key to Efficient Pervasive Computing", NSF Workshop series on Context-AwareMobile Database Management, Brown University, Providence, 24-25 January, 2002

[8] H. Rahman , R. Rahmani, and T. Kanter, "Realising Dynamism in MediaSense Publish/Subscribe Model for Logical-Clustering in Crowdsourcing", International Journal of Advanced Research in Artificial Intelligence (IJARAI), Vol. 3, No. 11, pp. 49-59, 2014

[9] M. Mamei, R. Menzes, R. Tolksdorf, F. Zambonelli, " Case studies for self-organization in computer science", Journal of System Architecture, pp. 443-460, Vol. 52 (2006), Issues 8-9, Elsevier

[10] Wikipedia, "Self-Organization", http://en.wikipedia.org/wiki/Self-organization [Last Accessed: February 2015]

[11] F. Dressler, "Self-Organization in Massively Distributed Systems – Methods and Techniques"

[12] M. A. Razzaque, S. Dobson, and P. Nixon, "Enhancement of Self-organization in Wireless Networking through a Cross-layer Approach," First Int'l Conf. ADHOCNETS, 2009.

[13] Kephart, J. O. (2005, May). Research challenges of autonomic computing. In Software Engineering, 2005. ICSE 2005. Proceedings. 27th International Conference on (pp. 15-22). IEEE.

[14] F. Zambonelli, O.F. Rana, "Self-Organization in Distributed Systems Engineering: Introduction to Special Issue", IEEE Transactions on Systems, Man, and Cybernetics,Vol. 35, No. 3, 2005

[15] R. Rahmani, H. Rahman, and T. Kanter, "Context-Based Logical Clustering of Flow-Sensors - Exploiting HyperFlow and Hierarchical DHTs", In Proceeding(s) of 4th International Conference on Next Generation Information Technology, 2013 ICNIT, June 2013

[16] T. Kanter et al., "MediaSense | The Internet of Things Platform", http://www.mediasense.se/ [Last Accessed: 08-February-2014]

[17] H. Rahman , R. Rahmani, and T. Kanter, "Enabling Scalable Publish/Subscribe for Logical-Clustering in Crowdsourcing via MediaSense", IEEE Science and Information Conference 2014, August 27-29, 2014, London, UK

Data Center Governance Information Security Compliance Assessment Based on the Cobit Framewok

(Case Study The Sleman Regency Data Center)

Andrey Ferriyan
Dept. Of Computer Science
Gadjah Mada University
Yogyakarta, Indonesia

Jazi Eko Istiyanto
Dept. Of Computer Science
Gadjah Mada University
Yogyakarta, Indonesia

***Abstract*—One of control domain of Cobit describes information security lies in Deliver and Support (DS) on DS5 Ensure Systems Security. This domain describes what things should be done by an organization to preserve and maintain the integrity of the information assets of IT where this all requires a security management process. One of the processes is to perform security monitoring by conducting periodic vulnerability assessment to identify weaknesses. Because Cobit is not explained technically, so it needs a method to utilize data that has been standardized. One of the standardized databases for vulnerability is CVE (Common Vulnerabilites and Exposures).This study aims to assess current condition of Data Center on Department of Transportation, Communication and Information Technology at Sleman Regency and assess the maturity level of security as well as providing solutions in particular on IT security. Next goal is to perform vulnerability assessment to find out which are the parts of the data center that may be vulnerable. Knowing weaknesses can help evaluate and provide solutions for better future. Result from this research is to create tool for vulnerability assessment and tool to calculate maturity model.**

Keywords—COBIT; CVE; maturity model

I. INTRODUCTION

Department of Transportation, Communication and Information Technology at Sleman Regency is one of the agencies that have the function of providing construction administration, development and management of communications network infrastructure. The agency has responsibility for managing the communications network infrastructure. In the development the infrastructure there are several incidents that have occurred. Several subdomain have been defaced by cracker. Distributed Denial Of Service (DDOS) attacking VoIP server. Remote security hole take place in the server where management authority not from Departemen of Transportation, Communication and Information Technology but the server itself lies on Data Center at Sleman Regency.

The evaluations from incidents conducted merely when there is a problem and the agency don't have evaluation planning and concept in safety evaluation in accordance with standards.

According to [1], to obtain a comprehensive security of the system, it is important to do the assessment and evaluate all aspects of the start of computer network security, application security, operating system security, database security, physical security and the environment.

Standard that can be used to assess the condition of governance data center is using the framework called Control Objectives for Information and Related Technology (COBIT). Today the use of Cobit framework is pretty much widely adopted and used as one of the standards in conducting research on the assets associated with information technology.

An application which can help an auditor is needed because not all applications can be applied in governance. Therefore, it needs to make an application that can be used to help a security auditor to evaluate the security of government information.

This paper describes how to assessing the data center using two methods. First by checking the compliance based on the Cobit framework and second by vulnerability assessment using tool combine with vulnerability standardized database.

II. LITERATURE REVIEW

Many have conducting research by presenting the results from the Cobit framework. However, Cobit to the security or vulnerability testing has been no detailed assessment. It's because Cobit is process oriented and works on management level. Comprehensive study conducted by comparing the level of maturity of the level of management. Cases studied tended to focus on one area, the management level or just the technical side.

According to [2], organization's management of IT governance requires the application of information technology environment to measure existing and planned in advance. Focus research only on management side and not technical issues where this is not much different from what is done by the [3] in his study.

According [3], vulnerabilities occur in the object under study is more to human error as an error in the conduct of data input, data duplication, deletion is done illegally, the absence of data storage settings, the absence of a recovery strategy the

data in the event of damage and others. The focus of research is over the control of the management and assessment of existing conditions.

Contrast to [3] where the focus of their research is more to the technical side. It mentions that the weakness of the system is classified into several pieces which are application vulnerabilities, network vulnerabilities, and host vulnerabilities. Research can not be considered comprehensive because it is merely checking individual.

Research conducted by [4] can help the system administrator to find configuration errors and then will be adjusted so that there is an error can be corrected. However, there is still a shortage of which is a new environment can be checked only one type only the server-based server Apache, PHP, MySQL.

As [5] who did research on the classification of network vulnerabilities with a research focus on the signs of the attack resulting from a firewall or IDS devices. Measuring threat approach to estimate the effects of attacks that occur in a computer network. It's utilizing the Common Vulnerability Scoring System (CVSS) to measure the amount of threat generated in the event of an attack.

III. Research Methods

Steps in research can be seen in Figure 1.

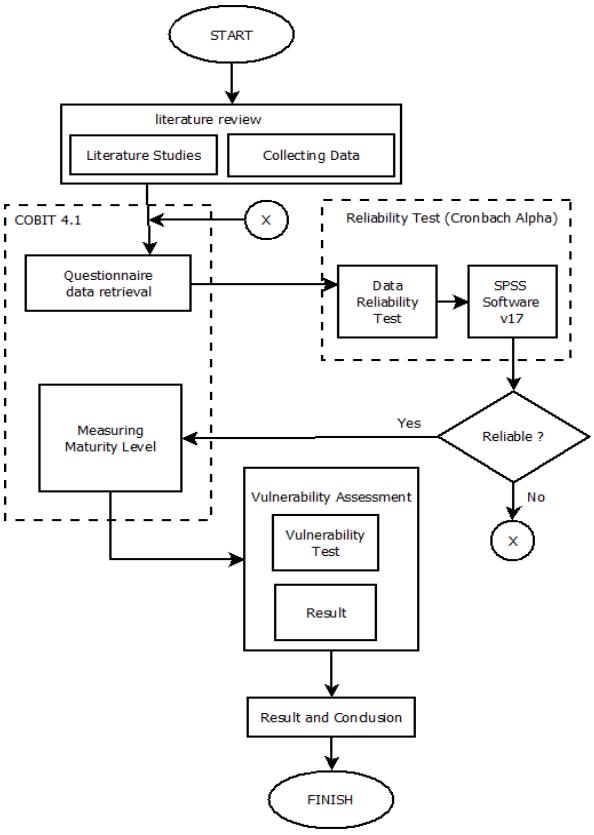


Fig. 1. Research steps

There are two methods conducting this research as described before. First by checking the compliance using Cobit. Cobit has a method to calculate maturity model how far the data center by looking at the security management process. In what level security management has been running well. Cobit will calculate it from level 0 to level 5. Of course this doesn't mean level 5 is very secure but by following it and using proper procedure we are convinced that the security management process run in accordance with the applicable procedures as stated by Cobit framework.

A. Literature Review

Studying literatur review of research related to Cobit and vulnerability assessment. Data collection was conducted by collecting technical documentation related to the data center like IP addresses, type of operating system, UNIX-like or Windows.

B. Questionnaire and Data Reliability

Questionnaires were distributed to the staff of the department. The questionnaire was taken from COBIT 4.1 Deliver and Support (DS) 5 Ensure Systems Security. Before the result from questionnaire can be used, its necessary to test the reliability first. Testing data reliability using Cronbach Alpha formula. Cronbach Alpha using scale from 0 to 1. If result from the formula equal or more than 0.6 then data considered reliable. Result from reliable data can be used to calculate maturity model. Based on the analysis, a tool for calculate of maturity model and a tool for vulnerability assessment are needed.

Vulnerability assessment tool will use CVE as standardized vulnerability database. The tool will use binary parsing to parse binary file within servers. If the version from binary version match from CVE list then it is considered vulner.

C. Implementation

Implementation from maturity model tool consists of :

- Reliability test to determine the level of reliability of measurement before continuing on maturity model calculations.
- Calculation of maturity analysis generated after the existing questionnaires obtained from the respondents.
- Reliability test to determine the level of reliability of measurement before continuing on maturity model calculations.

Implementation from a vulnerability tool consists of :

- Collecting binary version from all of servers.
- Comparing between binary version and CVE list. If matched then its considered vulner.

IV. Result and Discussion

After the tool is made, tests were done on two aspects, first the maturity of model-based testing, second testing from six servers that reside in the data center.

A. Maturity Model Testing

Maturity model test at the Department of Transportation, Communication and Information can be seen in Table 1

TABLE I. RESULTS OF MATURITY MODEL TESTING

Maturity Level	**All Questions**	**Total Question**	**Maturity Value**	Maturity Value Normalization	Maturity Model
Non Existent	5	7.5	1.5	0.044	0
Initial	6	35.5	5.917	0.175	0.175
Repeatable	8	60	7.5	0.222	0.444
Defined Process	7	46.5	6.643	0.197	0.591
Managed and Measureable	12	66	5.5	0.163	0.652
Optimised	10	67	6.7	0.198	0.99
Total			33.76	1	**2.852**

Table 1 shows that the value of the maturity model 2,852 value means the value at level 2.

B. Testing Vulnerability On Six Servers

The test is performed to determine the extent of the servers in the data center experienced technical vulnerabilities. This can be shown in Table II.

TABLE II. THE RESULT OF TESTING VULNERABILITY

No	Server Name	Severity Low	Severity Medium	Severity High	Total Vulnerability
1	Slemankab.go.id	3	0	2	5
2	Subdomain Slemankab	1	0	0	1
3	Web Perijinan	3	0	2	5
4	Database Perijinan	3	0	2	5
5	Web LPSE	0	1	1	2
6	Database LPSE	2	3	2	7
	Total	12	4	9	25

V. CONCLUSION AND SUGGESTIAN

A. Conclusion

Based on the test results that have been obtained, it can be concluded:

- The result of the questionnaire maturity model calculations COBIT 4.1.
- Maturity Model from Deliver and Support domain 5 shows the maturity value of the model is 2.852 for the Department of Transportation, Communication and Information.
- The results of the model calculation of maturity levels reached by the Department of Transportation, Communication and Information Technology is a level 2 or Repeatable for current conditions.
- Tools are made to calculate the maturity model has been proved correct by manual calculation using the formula in a spreadsheet.
- Tools are made to perform security testing failed to detect the presence of several vulnerabilities found in servers are tested.

B. Suggestion

The research using two methods has limitations that can be used as a reference for future development, suggested few things:

1) As for method one, for further testing maturity model involves calculating the expected maturity attributes such as awareness and communication, policy standards and 6 procedures, and automation tools, skills and expertise, responsibility and accountability and goal setting and measurement .

2) The necessity of making plans related to IT security and the solution where these plans appear based on the analysis of existing risks.

3) Scheduled for reporting security either in the form of a log or chart that can be read in conjunction with the existing security conditions both recent and in the distant past .

4) As for method two, PyCVE tool or script that is used in the research is still need to improve because not all application recorded on listing_file.txt readable version due to limitations in parsing binary file.

5) The tool for vulnerability assessment need more accurate in mapping for easier find any applications that may be vulnerable to later do an update on the server

REFERENCES

[1] Sayana, S., A., 2003, Approach to Auditing Network Security, Information Systems Control Journal Volume 5

[2] Lainhart IV, J., W., 2000, CobiT : A Methodology for Managing and Controlling Information and Information Technology Risks and Vulnerabilities, Journal Of Information Systems

[3] Pribadi, Y., I., 2011, Penilaian Kondisi Kini Tata Kelola Data Kependudukan Pada Aspek Pengelolaan Data Dengan Menggunakan Kerangka Kerja COBIT (Studi Kasus Kota Pontianak), Tesis, Jurusan Ilmu Komputer FMIPA, UGM, Yogyakarta.

[4] Eshete, B., Villafiorita, A., and Weldemariam, K., 2011, Early Detection of Security Misconfiguration Vulnerabilities in Web Applications, In Proceedings of the 6th Conference on Availability, Reliability and Security (ARES2011), Vienna, Austria, 169-174

[5] Xi, R., Yun, X., Jin, S., dan Zhang, Y., 2011, Network Threat Assessment Based on Alert Verification, PDCAT 2011 Proceedings of the 2011 12th International Conference on Parallel and Distributed Computing, Applications and Technologies, Gwangju, Korea, 30-34

Intelligent Traffic Information System Based on Integration of Internet of Things and Agent Technology

Hasan Omar Al-Sakran
Management Information Systems Department
King Saud University
Riyadh, Saudi Arabia

***Abstract*—In recent years popularity of private cars is getting urban traffic more and more crowded. As result traffic is becoming one of important problems in big cities in all over the world. Some of the traffic concerns are congestions and accidents which have caused a huge waste of time, property damage and environmental pollution. This research paper presents a novel intelligent traffic administration system, based on Internet of Things, which is featured by low cost, high scalability, high compatibility, easy to upgrade, to replace traditional traffic management system and the proposed system can improve road traffic tremendously. The Internet of Things is based on the Internet, network wireless sensing and detection technologies to realize the intelligent recognition on the tagged traffic object, tracking, monitoring, managing and processed automatically. The paper proposes an architecture that integrates internet of things with agent technology into a single platform where the agent technology handles effective communication and interfaces among a large number of heterogeneous highly distributed, and decentralized devices within the IoT. The architecture introduces the use of an active radio-frequency identification (RFID), wireless sensor technologies, object ad-hoc networking, and Internet-based information systems in which tagged traffic objects can be automatically represented, tracked, and queried over a network. This research presents an overview of a framework distributed traffic simulation model within NetLogo, an agent-based environment, for IoT traffic monitoring system using mobile agent technology.**

Keywords—Intelligent Traffic; Internet-of-Things; RFID; Wireless Sensor Networks; Agent Technology

I. INTRODUCTION

In recent years popularity of private motor vehicles is getting urban traffic more and more crowded. As result traffic monitoring is becoming one of important problems in big smart-city infrastructure all over the world. Some of these concerns are traffic congestion and accidents that usually cause a significant waste of time, property damage and environmental pollution. Any type of congestion on roads ultimately leads to financial losses. Therefore, there is an urgent need to improve traffic management. The appearance of the Internet of Things (IoT) provides a new trend for intelligent traffic development.

This research proposes to employ the IoT, agent and other technologies to improve traffic conditions and relieve the traffic pressure. Information generated by traffic IoT and collected on all roads can be presented to travelers and other users. Through collected real-time traffic data, the system can recognize current traffic operation, traffic flow conditions and can predict the future traffic flow. The system may issue some latest real-time traffic information that helps drivers choosing optimal routes. Therefore, the system can precisely administrate, monitor and control moving vehicles. Constructing an intelligent traffic system based on IoT has a number of benefits such improvement of traffic conditions, reduction the traffic jam and management costs, high reliability, traffic safety and independence of weather conditions [1, 2].

Such traffic IoT must include every element of traffic such as roads, bridges, tunnels, traffic signals, vehicles, and even drivers. All these items will be connected to the internet for convenient identification and management through sensor devices, such as RFID devices, infrared sensors, global positioning systems, laser scanners, etc.

Traffic IoT provides traffic information collection and integration, supporting processing and analysis of all categories of traffic information on roads in a large area automatically and intelligently. Thus, modern traffic management is evolving into an intelligent transport system based on IoT.

Traffic requires suitable information about services and logistics available on the road and therefore the system can become more self-reliable and intelligent. With a number of WSN and Sensor enabled communications, an IoT of data traffic will be generated. This traffic monitoring applications need to be protected to prevent any security attack frequent in urban cities. Few such prototypes implementations can be found in [3, 4] and the Smart Santander EU project [5].

The aim of this paper is to present a framework for real-time traffic information acquisition and monitoring architecture based on the IoT utilizing wireless communications. The primary characteristic of the proposed traffic information infrastructure is its capability of integrating different technologies with the existing communication infrastructures. The proposed architecture allows gathering real-time traffic data generated by sensory units and monitoring the traffic flow using multi-agent based system. Agents can perform specific tasks with a degree of intelligence and autonomy, and interact with their environment in a useful way without human intervention thus decreasing network load, facilitating heterogeneous IoT devices, providing support for collaboration and interoperability in IoT and programmable RFID and WSN,

overcoming network latency, and asynchronous and autonomous execution.

The remainder of the paper is organized as follows. Background on IoT is discussed in section 2. Section 3 presents related work. Framework structure of the proposed traffic system is introduced in section 4. Section 5 describes the agent-based approach for the development of intelligent traffic information system. Discussion of the proposed traffic simulation framework is presented in section 6. Finally, section 7 is devoted to conclusions and future work.

II. INTERNET OF THINGS

During past few years recent communication paradigm - the internet of things - has gained significant attention in academia as well as in industry because it represents an enormous opportunity for cost savings and new revenue generation across a wide range of industries. The main reasons behind this interest are its capabilities. IoT can be used to create a world where all smart objects of our everyday life are connected to the Internet and interact with each other with minimum human involvement to reach a common goal [8]. The term Internet of Things was first appeared by Kevin Ashton [9] in the context of supply chain management.

Gartner forecasts that the IoT will reach 26 billion units by 2020, up from 900 million just five years ago, and this will impact the information available to supply chain leaders. According to Cisco's study, cities all over the world are to claim $1.9 trillion in value from IoT over the next decade by building smarter cities based on smarter infrastructure, through providing optimal traffic management, parking, and transit services [10].

The enabling technologies that are expected to form the building blocks of the sensing and communication technologies in IoT are Wireless Sensor Networks (WSN) and RFID-based networks connected together through the Internet or other technologies and protocols. RFID is considered as one of the leading technologies mainly due to its low cost, and its strong support from the business community. RFID can transform everyday objects into smart objects. Sensor network integrates different technologies, such as sensor, distributed information processing, embedded computing and wireless communications. Sensors and RFID are playing a significant role in constructing IoT. Multiple RFID and sensors with computing and communication power are connected into wireless networks and cooperate with each other to exchange collected data with the physical world to accomplish specific tasks.

Implementation of IoT relies on the integration of RFID systems, WSNs, and intelligent technologies. RFID and wireless data communication technology are used to construct a network which covers everything. Objects such as RFID tags and readers, sensors, actuators, mobile phones, smart devices, embedded computers, etc., will be included into the network and will interact with each other through unique addressing schemes [11]. These objects have actuating, processing, storing and networking capabilities. With the advances in sensor technology, sensors will be embedded within all the objects around us. The result will be the generation of huge amounts of data which will have to be stored, processed and presented in efficient and easily interpretable form. IoT allows people and various objects to be connected anytime and anywhere with anything and to any service, and use any network; and communicate with each other in real time as long as they are online [12, 13].

Other necessary components include cloud, data modeling, storing, processing, and communication technologies [14]. The major wireless technologies used to build wireless sensor networks are wireless personal area network (Bluetooth), wireless local area network (Wi-Fi), wireless metropolitan area network (WiMAX), wireless wide area network (3G/4G mobile networks) and satellite network (GPS). A typical structure of a RFID based sensor network is presented in fig. 1. It consists of wireless low-end RFID sensor nodes that generate data (tags) and high-end RFID sensor nodes that retrieving data from the low nodes. Data collected by the high nodes are sent to mobile static nodes (readers). Readers send the data to wireless low-end computational devices (base stations). These devices perform a certain amount of processing on the sensor data. Then data sent to high-end computational servers through the internet (or other network) to be processed further and there data will be shared and stored.

III. RELATED WORK

A number of researchers have dealt with the problem of intelligent traffic monitoring and controlling, and as a result of their efforts several different approaches have been developed. Pang et al. [15] proposed a traffic flow prediction mechanism based on a fuzzy neural network model in chaotic traffic flow time series. Bhadra et al. [16] applied agent-based fuzzy logic technology for traffic control situations involving multiple approaches and vehicle movements. In [17] the authors developed strategies to integrate different dynamic data into Intelligent Transportation Systems. Patrik et al. [18] proposed a service-oriented architecture (SOA) for an effective integration of IoT in enterprise services.

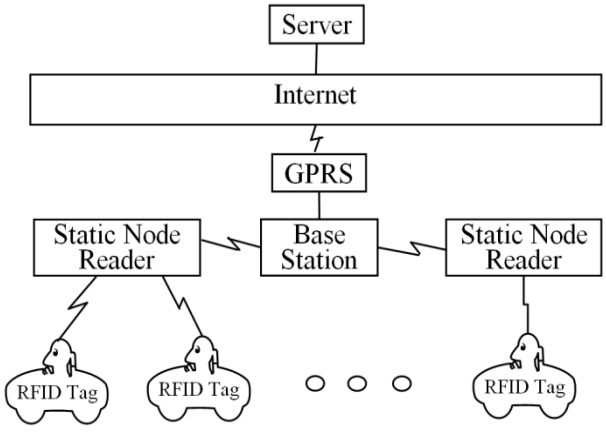


Fig. 1. RFID-based sensor network

Recently researchers shifted their attention to revolutionizing paradigm of the Internet of Things, which resulted in constructing of a more convenient environment composed of various intelligent systems in different domains such as intelligence business inventories, health care,

intelligent home, smart environment, smart metering, supply chain logistics, retail, smart agriculture, monitoring electrical equipment, etc. [19-22], while it is still in the early stage in case of intelligent transportation system with respect to their needs [23-26]. Different IoT systems such as UbiComp [27], FeDNet [28, 29] are using message simple passing techniques for communication. Such techniques consume a large amount of bandwidth and energy.

Agent technology has been implemented in different aspects of the traffic systems such as handling traffic congestion by monitoring the current traffic congestion and providing the optimal route for a vehicle [30-32]. Fortino et al. [33] proposed an architecture integrating agents and cloud computing to develop decentralized smart objects within IoT, while Godfrey et al. [34] used mobile agent to handle not just the communications among devices within the IoT but to conduct searching for needed resources.

IV. Framework Structure of Proposed System

The major tasks of the proposed system are detecting mobile objects and their location, identifying mobile objects and transmitting acquired data to the monitoring and controlling center for processing.

A general overview of the proposed intelligent traffic system is shown in Table 1. The structure of the proposed traffic IoT system consists of three layers: application, network and acquisition.

Main functions of the application layer are collecting, storing, and processing traffic data to produce value-added services; presenting the interface of traffic IoT to users and analyzing received information from acquisition layer according to the different needs.

The application layer includes the following subsystems:

- Intelligent Driver Management Subsystem: drivers can acquire real-time traffic information with minimum delay.
- Vehicle Guidance and Road Information Management Subsystem: monitoring number of vehicle on one road, tracking vehicle's violation, sending warning massages, guide drivers to avoid possible crowded sections based on the prediction of the traffic network, real-time traffic navigation, etc.
- Intelligent Traffic Management Subsystem: the traffic system database contains data from vehicle sensors, weather information from environmental sensors, and information on traffic flows. The subsystem processes received information and shares it through the interface with other subsystems. It allows tracing the location of a vehicle fast and accurate and optimizing traffic scheduling.

TABLE I. Intelligent Traffic IoT

Application Layer	Intelligent Traffic Management	Intelligent Driver Management	Information Collection & Monitoring	Information Services
Network Layer	Internet	WiFi, 3G/4G	WiMax	GPS, GPRS
Acquisition Layer	RFID	RFID Reader	WSN	Intelligent Terminals

- Information Collection And Monitoring Subsystem: real-time distribution the information of road conditions, weather information, accident monitoring, etc. The subsystem merges data from different subsystems and provides it to end users in a suitable format.
- Information Service Subsystem: performs online vehicle information query and dynamic statistic analysis of real-time traffic flow, tracks a specific vehicle and generates reports for traffic management department.

The network layer, also called transport layer, is constituted by all sorts of private networks, Internet, wired and wireless communication networks, network management system, global positioning system(GPS), wireless general packet radio service (GPRS), worldwide interoperability for microwave access (WiMax), wireless fidelity (WiFi), Ethernet, and corporate private networks. It is responsible for transmitting data with high reliability and security, and processing the information coming from acquisition layer. GPRS provides high-speed wireless IP services for mobile users and fully supports the TCP/IP. The wireless communication channels used by the devices may include any of the prevailing standards such as IEEE 802.11, Zigbee or Bluetooth, etc.

Acquisition layer is constituted by all kinds of sensors and sensor gateways such as RFID, WSN, cameras, intelligent terminals to transmit data of mobile objects and other sensors used to collect real-time traffic and object identification information. It serves as a source of all types of information (for example, identified objects, traffic flow, etc.) collected from the physical world. Its main functions are to collect real-time information from IoT sensors, monitor objects and transfer data to the network layer.

The system utilizes wireless sensors to obtain real-time traffic information, such as traffic condition on each road, number of vehicles, average speed, and so forth. Utilization of wireless sensors is very appropriate due to their low power consumption, low cost, distributed processing and self-organization. In order to achieve large-scale network layout the system uses wireless cluster sensor network. Each cluster has a set of wireless sensors and each set is represented by the head node. Data at the head nodes are delivered to the backend system by a mobile agent.

Already some new vehicles are equipped with GPS and sensors capable of receiving and sending driving information to the monitor and control center via the satellite communication facilities at any time. GPS could be connected with the wireless sensor networks which can be used for measuring speed, driving direction.

V. Development of an Agent-Based Intelligent Traffic Information System

There are a large number of heterogonous devices within the traffic monitoring system using IoT. Among challenges of full deployment IoT is making complete interoperability of these heterogeneous interconnected devices which require adaptation and autonomous behavior. The major issue in IoT is the interoperability between different standards, data formats, heterogeneous hardware, protocols, resources types, software and database systems [35, 36]. Another issue is necessity of an intelligent interface and access to various services and applications. It seems that mobile agents are a convenient tool to handle these issues, provide means for communication among such devices and handle the IoT interoperability. Adding to that mobile agent is a perfect choice in cases of disconnection or low bandwidth, passing messages across networks to undefined destination and to handle the interoperability of IoT. All messaging exchanges among agents are established via the TCP/IP Protocol.

A software agent is an autonomous executable entity that observes and acts upon an environment and acts to achieve predefined goals. Agents can travel among networked devices carrying their data and execution states, and must be able to communicate with other agents or human users. A multi-agent system is a collection of such entities, collaborating among themselves with some degree of independence or autonomy.

Applying agent technology in the process of monitoring and control traffic is new approach. Such technology perfectly fits for distributed and dislocated systems like traffic monitoring and controlling due to its autonomy, flexibility, configurability and scalability thus reducing the network load and overcoming network latency. Agents can also be used to pass messages across networks where the address of destination traffic device is unidentified. Each traffic object is represented as a software agent (an intelligent object agent). In this infrastructure the extremely large variety of devices will get interconnected, and will be represented by its own intelligent agent that collects information and responds to others' requests. Agents will provide their functionality as a service. Autonomous intelligent agents are deployed to provide services necessary for the execution of functional tasks in each layer of the proposed architecture.

An agent is embedded within each device and each device supports all agent functions such as migration, execution. Whole system can be controlled by the specific application written for each device's mobile agent defining how it should behave and act intelligently. Mobile agents within the network migrate from one node to another allowing the devices to pass information to others, retrieve information and discover available resources.

Main IoT Traffic agents are:

- Traffic Mobile Agent: Transmits/receives different types of information to/from other objects the Internet; interprets the data coming from other objects (RFID, sensors, users), and provides a unified view of the context; communicates with other agents in the network to accomplish a specific task. All messages sent from this agent will be transferred to the traffic management system and communicate directly with a static agent of the intended application of the traffic management system mentioned above.
- User Agent: provides users with real-time information of entities residing in the system. The user agent is a static agent that interacts with the user. It is expected to coordinate with mobile agents.
- Monitor Agent: monitors the system to detect contingency situations and triggers some actions to react to some tag reading events on behalf of a smart traffic object, for example in emergency cases.
- RFID Agent: responsible for reading or writing RFID tags. When reading a tag, according to the data retrieved from it, this agent performs appropriate operations in handling a single task on behalf of a smart object of the associated RFID and to migrate to different platforms at run time.
- Sensor Agent: receives, processes data that have been read from the associated sensor and saves (or send it somewhere).
- Traffic Light Agent: detects irregular traffic conditions and changes the traffic control instructions right away.
- Camera Agent: is responsible for image collecting. All communications between camera agent and video Web server are conducted via the network layer. Camera agent can takes advantage of the existing infrastructure of the camera-based traffic monitoring systems that already available in many cities.

The traditional traffic monitoring system based on image-processing technology has many limitations. One of them is the impact of the weather. In case of thick dust, heavy rain, etc., the license plate cannot be seen clearly, so its image cannot be captured. The development of e-plate based on RFID provides a good opportunity for intelligent traffic monitoring and vehicle's identification and tracking [37]. If no agents are associated with the RFID tags (identification-centric RFID systems), then they may function as an independent set of programs for tag processing and communicate using standardized software agent protocols. The author suggests utilizing the agent technology within the e-plate based on RFID and other traffic objects to fully realize the combined potential of RFID and software agent technology.

An RFID-based smart traffic object (code-centric RFID systems) requires a substantial amount of memory space to store traffic object logics and data. The code-centric RFID systems can be used to store a mobile agent into the RFID tags that will enable integration with other parts of the traffic system. Using such technology in the Traffic Information System will eliminate the need for searching of the associated

RFID-code information from a database and reduce overall system response time by retrieving service information from the tags [38], thus achieve faster service responses and perform on-demand actions for different objects in different situations. Each smart vehicle's RFID object consists of two components, namely, object processing logics and object data [39]. The object data contains a global unique Electronic Product Code (EPC) code as its unique identifier. Each RFID-tagged traffic object may be assigned an IPv6 Mapped EPC address [40]. The IoT networks are expected to include billions of devices, and each shall be uniquely identified. A solution to this problem is offered by the IPv6, which provides a larger address space of 128-bit address field to accommodate the increasing number of devices in IoT, thus making it possible to assign a unique IPv6 address to any possible device in the IoT network.

RFID can be used as a transponder in vehicle registration plate equipped with a RFID tag and sensors so that each car can get data it needs from the spot and deliver to assigned destination. The vehicle RFID tag stores information on the vehicle and its owner, such as plate number, vehicle type, speed, time when the car reaches the monitoring point, driver's name and license number. It can be used to estimate the number of vehicles in the road, average speed of vehicles, vehicle density, etc. The data from each vehicle is captured by fixed or mobile RFID reader at a monitoring station as information of the vehicle and will be sent to central server unit for collecting, processing and storing. Once system connects to the internet, all information of vehicles on each road segment is immediately saved in database and can be used for any purpose and application (vehicle tracking, monitoring or traffic information, etc.).

When a vehicle with an RFID tag passes through each monitoring station along the road, the RFID reader at those points will automatically read the tag data related to the vehicle and its owner and transmit to the wireless sensor active nodes. These nodes send accumulated data to the cluster head node. At the same time, a GPS receiver installed at the monitoring station can communicate with GPS satellites to obtain its position information that is taken as a position parameter of the vehicle. Then the data is transmitted using GPRS scheme to the real-time central database where the data is constantly updated to ensure data reliability.

VI. Traffic Simulation Framework

To justify the proposed system online distributed traffic simulation was conducted. Simulation allows us to observe the properties, characteristics and behaviors of the traffic system. Based on detailed real-time data collected from the distributed online simulations, the IoT traffic system can provide accurate information necessary for near real-time traffic decisions.

The whole traffic IoT network is partitioned into dynamic overlapped sections, and a simulation processor is mapped to each section. Each simulation will be supplied with real-time data from nearby RFIDs and sensors and enabled to run continuously. The overall distributed simulation consists of a collection of such segment simulations where each small segment of the overall traffic IoT network is modeled based on local criteria. Each simulation segment is operating in an asynchronous mode, meaning each simulator executes independently of other simulators and the simulation server.

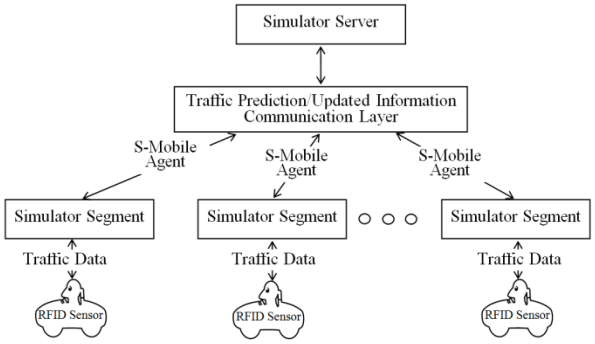


Fig. 2. Distributed online traffic simulation framework

These simulation segments are allowed to exchange information on vehicles moving from one simulation segment to another. Each simulator's segment locally models current traffic conditions and concentrating only on its area of concern. A simulator's segment, for example, might model some set of roads and intersections of that segment, and predict the rates of vehicle flow on links carrying vehicles out of that segment. Each segment shares its predictions with other simulation segments to create an aggregated view of both the individual segment's area of interest and the overall of traffic system. Simulators' segments publish their current traffic state information (speed, travel time, flow rate, etc.) and their predictions to the simulation server. An aggregation of all simulation segments provides an accurate estimation of a future state of the system.

The general model of distributed traffic simulation framework described in fig. 2. The simulation server disseminates information among the simulator segments, coordinates all simulators' segments and provides a predictive model of traffic conditions in specified traffic areas by analyzing and integrating the results of distributed simulators of those areas. The simulation server maintains state information of current and future operations of the traffic network such as flow rates, average speed, and the time when that information was generated. Running online simulations are integrated with traffic information system infrastructure to receive real-time traffic data and this overall simulation provides detailed information required for prediction of the system future states of the system. Detailed traffic information (such as speed, location, average acceleration of vehicles on the network segment and the current state of traffic control devices) generated during simulation is saved and managed on the simulation server.

Online distributed traffic simulation is a powerful approach for analyzing the characteristics and behavior of the traffic system and determining traffic conditions and help to reduce vehicle delay time of on the road, traffic congestion without the need of making costly changes in real world; prevent dangerous situations and delays by broadcasting messages informing drivers in the area to avoid congested roads [41]. It will be beneficial to transportation management as well as urban planning and architecture working on enhancement of

the roads capacity, building new roads or to improve the existing roads and improvement of public transportation systems.

The current large-scale distributed simulation methodologies require tremendous network bandwidth and huge amount of computation by each simulator host. Mobile agents are used to reduce the communications loads placed in the network. Agents communicate with a specific simulation segment, providing all of the state information that was sent to the simulator server.

NetLogo simulator has been used for modeling a collection of adjacent intersections. Static and mobile agents represent different features of the network. Motor vehicles have been modeled individually within NetLogo using mobile agents. Simulation can be run on several computers. NetLogo allows giving instructions to large number of independent agents which could all operate at the same time. In this cause the NetLogo model runs in a single machine computing environment, but it can be extended to run on cluster of computers.

Four types of agents used in NetLogo: patches are used to represent static agent, turtles for mobile agent; links are used to make connections between turtles; and the observer for observing everything going on in the simulated environment [42]. The environment of the NetLogo is written entirely in Java, therefore patches and turtles are programmable by the user of NetLogo in Java language. In this simulation, the agent entities are vehicle, traffic lights, and sensors of intersections and lanes. Agents are created and randomly distributed over the network of intersections. A random number of vehicles were generated according to limits defined in the model. Sensors obtained the number of passing vehicles. The traffic lights action are based on goals of minimizing the waiting time of vehicles travelling through intersections and increasing throughput of vehicles that successfully pass through these intersections.

During each run, the following indicators were produced: not moving vehicles, average waiting time and average of speed of the vehicles in a time step. The human factors play an important role in traffic systems. In most cases the driver's behavior are unpredictable. Modeling of drivers' behavior using agent-based has been performed based on techniques proposed by [43].

The simulation has 'setup' and 'go' buttons. The 'setup' button calls a procedure to reset the model to the initialization state, and the 'go' button calls a procedure that carries out all actions for each simulation run. All visual aspects are managed by the NetLogo simulation, and after every run visualizations are automatically updated. The interface and performance evaluation of the simulation results are shown in fig. 3.

VII. Conclusions and Future Research

This paper presents a real-time traffic information collection and monitoring system architecture to solve the problem of real-time monitoring and controlling road vehicles.

The proposed architecture employs key technologies: Internet of Things, RFID, wireless sensor network (WSN), GPS, cloud computing, agent and other advanced technologies to collect, store, manage and supervise traffic information.

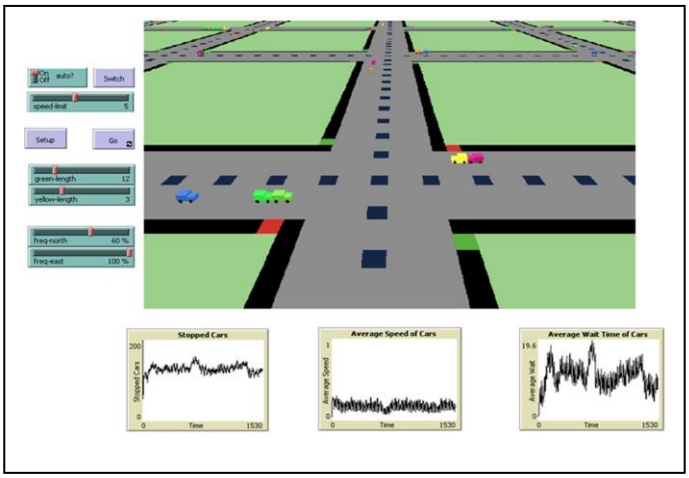


Fig. 3. Interface and performance evaluation of the simulation results

Agents provide an effective mechanism for communication amongst networked heterogeneous devices within the traffic information system.

The proposed system can provide a new way of monitoring traffic flow that helps to improve traffic conditions and resource utilization. In addition, transport administration department, using real-time traffic monitoring information, can in time detect potentially dangerous situations and take necessary actions to prevent traffic congestion and minimize number of accidents thus ensuring safety of road traffic. In general, the IoT will play an important role in the traffic management enhancing the efficiency of information transmission, improving traffic conditions and management efficiency, traffic safety, and reducing management costs.

However, the proposed traffic system based on the IoT consists of a large number of RFIDs and sensors that transmit data wirelessly. This calls for improved security to protect such massive amounts of data and privacy of users. It's a challenge for future research to ensure the security of smart objects in the traffic monitoring management system in case of a cyber-attack or an intentional interest to a member of the IoT infrastructure. IoT requires modification of network connectivity models and readiness for massive increase in amount of real-time information. To achieve that, interaction communication models must be redesigned to include machine to machine and people to machine communications. Another research area is processing and analytics of large volumes of disparate data from Traffic IoT system to create applications that improve the flow of vehicles throughout the city.

References

[1] Laisheng Xiao, "Internet of Things: a New Application for Intelligent Traffic Monitoring System", Journal of Networks, 2011, vol. 6, No. 6.

[2] J. R.Molina, J. F. Martínez, P. Castillejo and L. López, "Combining Wireless Sensor Networks and Semantic Middleware for an Internet of Things-Based Sportsman/Woman Monitoring Application", Sensors, 2013, vol. 13, pp. 1787-1835.

[3] European Lighthouse Integrated Project - 7th Framework, Internet of Things - Architecture. http://www.iot-a.eu/ , 2012.

[4] K. Kotis, & A. Katasonov, "Semantic Interoperability on the Web of Things: The Smart Gateway Framework", In Proceedings of the Sixth International Conference on Complex, Intelligent, and Software Intensive Systems (CISIS-2012), Palermo, 2012.

[5] A. J. Jara, S. Varakliotis, A. F. Skarmeta and P. Kirstein, "Extending the Internet of Things to the Future Internet through IPv6 support", Mobile Information Systems, IOS Press, 2013.

[6] Laura Jarvenpaa, et al, "Mobile Agents for the Internet of Things", 17th International Conference on System Theory, Control and Computing (ICSTCC), 2013.

[7] Leppänen, Teemu, Liu, Meirong, et al., "Mobile Agents for Integration of Internet of Things and Wireless Sensor Networks", 2013 IEEE International Conference on Systems, Man, and Cybernetics.

[8] L. Atzori, A. Iera and G. Morabito, "The internet of things: A survey", Comput. Netw, 2010, vol. 54, no. 15, pp. 2787–2805.

[9] K. Ashton, "That Internet of Things thing", RFiD Journal, 2009.

[10] http://newsroom.cisco.com/release/1308288, 2014.

[11] D. Bandyopadhyay and J. Sen, "The internet of things - applications and challenges in technology and Standardization", Springer International Journal of Wireless Personal Communications, 2011, vol. 58, no. 1, pp. 49-69.

[12] P. Guillemin and P. Friess, "Internet of things strategic research roadmap", The Cluster of European Research Projects, Tech. Rep., September 2009, http://www.internet-of-things-research.eu/pdf/IoT Cluster Strategic Research Agenda 2009.pdf [Accessed on: 2011-08-15].

[13] D. Singh, G. Tripathi and A. J. Jara, "A survey of Internet-of-Things: Future Vision, Architecture, Challenges and Services", IEEE World of Forum on Internet of Things.

[14] B. Xu, "Key IOT Technology and Application Research", Applied Mechanics and Materials, 2014, vol. 543-547, pp. 3411-3414.

[15] M. Pang and X. Zhao, "Traffic Flow Prediction of Chaos Time Series by Using Subtractive Clustering for Fuzzy Neural Network Modelling," Proceedings 2nd International Symposium Information Technology Application, Washington – DC, 2008, pp. 23-27.

[16] S.Bhadra, A. Kundu and S. K. Guha, "An Agent based Efficient Traffic Framework using Fuzzy", Fourth International Conference on Advanced Computing & Communication Technologies, 2014.

[17] V.Katiyar, P. Kumar and N. Chand, "An Intelligent Transportation System Architecture using Wireless Sensor Network", International Journal Computer Applications, 2011,vol. 14, pp. 22-26.

[18] P. Spiess, S. Karnouskos, D. Guinard, D. Savio, O. Baecker, L. Souza, et al., "SOA-based integration of the internet of things in enterprise services", In: Proceedings of IEEE ICWS 2009, Los Angeles, pp. 1–8.

[19] Libelium Communications Distribuend as S.L., "50 sensor applications for a smarter world", 2014, available at http://www.libelium.com/top 50 iot sensor applications ranking/.

[20] D. Miorandi, S. Sicari, F. De Pellegrini and I. Chlamtac, "Internet of things: Vision, applications and research challenges," Ad Hoc Networks, 2012, vol. 10, no. 7, pp. 1497–1516.

[21] M. C. Domingo, "An overview of the internet of things for people with disabilities," Journal of Network and Computer Applications, 2012, vol. 35, no. 2, pp. 584–596.

[22] T.S. Lo´pez, D.C. Ranasinghe and M. H. Duncan McFarlane, "Adding sense to the Internet of Things An architecture framework for Smart Object systems", Personal Ubiquitous Computing, 2012, vol. 16, pp. 291–308.

[23] P. Pyykonen, J. Laitinen, J. Viitanen, P. Eloranta and Korhonen, "IoT for Intelligent Traffic System, IoT for intelligent traffic system", International Conference on Intelligent Computer Communication and Processing (ICCP), 2013 IEEE.

[24] C.Yulian, L. Wenfeng and J. Zhang, "Real-Time Traffic Information Collecting and, Monitoring System Based on the Internet of Things", 6th International Conference on Pervasive Computing and Applications (ICPCA), 2011 IEEE.

[25] X. Yu, F. Sun and X. Cheng, "Intelligent Urban Traffic Management System Based on Cloud Computing and Internet of Things", International Conference on Computer Science & Service System, 2012 IEEE, pp. 2169 – 2172.

[26] Y. Yin and J. Dalin, "Research and Application on Intelligent Parking Solution Based on Internet of Things", 5th International Conference on Intelligent Human-Machine Systems and Cybernetics (IHMSC), 2013, pp. 101 – 105.

[27] C. Goumopoulos and A. Kameas, "Smart Objects as Components of UbiComp Applications", International Journal of Multimedia and Ubiquitous Engineering, vol. 4, no. 3.

[28] F. Kawsar, "A Document based Framework for User Centric Smart Object Systems", Ph.D. Dissertation, Waseda University, Tokyo.

[29] F. Kawsar and T. Nakajima, "A Document Centric Framework for Building Distributed Smart Object Systems", in 2009 IEEE International Symposium on ObjectiComponentiService-Oriented Real-Time Distributed Computing, Tokyo, 17-20 March 2009, pp. 71-79.

[30] G. Nakamiti, V.E. Silva and J.H. Ventura, "An Agent-Based Simulation System for Traffic Control in the Brazilian Intelligent Cities Project Context", Proc. 2012 Agent Direct Simulation Conference, Orlando FL.

[31] B.Chen, H.H. Cheng, and J. Palen, "Integrating mobile agent technology with multi agent systems for distributed traffic detection and management systems", Transport Research, 2009, vol.17, no. 1, pp. 1-10.

[32] T. Karthikeyan and S.Sujatha, "Optimization of Traffic System using TCL Algorithm through FMSA and IMAC Agents", International Journal of Advanced Research in Computer Engineering & Technology (IJARCET), 2012, vol. 1, no 6.

[33] G.Fortino, A. Guerrieri, W. Russo and Cl. Savaglio, "Integration of Agent-based and Cloud Computing for the Smart Objects-oriented IoT", Proceedings of the 18th International Conference on Computer Supported Cooperative Work in Design, 2014 IEEE.

[34] W. W. Godfrey, S. S. Jha and B. N.Shivashankar, "On A Mobile Agent Framework for an Internet of Things", 2013 International Conference on Communication Systems and Network Technologies.

[35] A. Katasonov, O. Kaykova, et al.,"Smart Semantic Middleware for the Internet of Things," In: Proceedings of the 5th International Conference on Informatics in Control, Automation and Robotics, Intelligent Control Systems and Optimization, 2008, pp. 169-178.

[36] T.Leppanen, L.Meirong , et al., "Mobile Agents for Integration of Internet of Things and Wireless Sensor Networks", 2013 IEEE International Conference on Systems, Man, and Cybernetics (SMC), pp. 14 – 21.

[37] Evizal,T. Abd Rahman, S. K. Abdul Rahim, "RFID Vehicle Plate Number (e-Plate) for Tracking and Management System", 2013 EEE International Conference on Parallel and Distributed Systems, pp: 611 – 616.

[38] M. Chen, S. González, Q. Zhang and V. C.M. Leung, "Code-Centric RFID System Based on Software Agent Intelligence", IEEE Intelligent Systems, 2010.

[39] M. Chen, S. Gonzalez-Valenzuela, Q. Zhang and V. Leung, "Software agent-based intelligence for code-centric RFID Systems," IEEE Intelligent Systems, 2010, vol. 99.

[40] M. C. Chung, G. M. Lee, N. Crespi and C. C. Tseng, "RFID Object Tracking with IP Compatibility for the Internet of Things", 2012 IEEE International Conference on Green Computing and Communications, Conference on Internet of Things, and Conference on Cyber, Physical and Social Computing.

[41] W. Suh, M. P. Hunter and Richard Fujimoto, "Ad hoc distributed simulation for transportation system monitoring and near-term prediction", Journal of Simulation Modelling Practice and Theory, 2014, vol. 41, pp. 1–14.

[42] U. Wilensky, NetLogo User Manual version 4.1.3, Center for Connected Learning and Computer-Based Modelling, Northwestern University, Evanston, IL. Available at: http://ccl.northwestern.edu/netlogo/. As accessed on July 2014.

[43] R. J. F. Rossetti, R. H. Bordini and A. L. C. Bazzan, "Using BDI agents to improve driver modelling in a commuter scenario", Transportation Research, Part C, pp. 373–398, 2002.

Development of a Decision Support System for Handling Health Insurance Deduction

Shakiba Khademolqorani
Department of Industrial & Systems Engineering
Isfahan University of Technology
Isfahan, Iran

Ali Zeinal Hamadani
Department of Industrial & Systems Engineering
Isfahan University of Technology
Isfahan, Iran

***Abstract*—Effective hospital management involves such activities as monitoring the flow of medication, controlling treatment, and billing for the patient's treatment. A major challenge between insurance companies and hospitals lies in the way medical treatment expenses for insured patients are reimbursed. In some cases, the insurance deduction leads to the loss of revenues by hospitals. This paper proposes a framework for the handling insurance deduction that integrates three major methodologies: Decision Support Systems, Data Mining, and Multiple Criteria Decision Making. To exemplify the practical utility of the framework, it is used to study hospital services and insurance deductions are extracted from 200,000 documents in 150 hospitals in Iran. To classify the kinds of services, decision trees are developed to mine hidden rules in the data which are then modified on the basis of some performance measures. The rules are then extracted and ranked using the TOPSIS method. The results show that the proposed framework is capable of effectively providing objective and comprehensive assessments of insurance deductions.**

Keywords*—Hospital Management; Insurance Deduction; Decision Support Systems; Data Mining; Multiple Criteria Decision Making

I. INTRODUCTION

Hospital management involves a most complex decision making process that has to deal with huge arrangements related to such administrative and medical operations as identifying patients, processing healthcare benefits for the inpatient, supporting administrative functions, facilitating payments for the services, and assisting insurance providers in their quest for in-depth records of actual treatments provided. In such complex management systems, past medical histories, problems, demographics, laboratory data, and basic information are incorporated into one single system in order to accelerate clinical studies and drug administration to patients [1, 2].

In Iran, a plan was approved in 1985 for the autonomous management of hospitals, in which hospital costs are reimbursed from their own revenues. This made the financial management of hospitals more complicated than ever before. A majority of hospital revenues are reclaimed as per contracts with insurance companies which provide insurance policies to patients for hospital care and services [3].

A big challenge facing hospital managers is their transactions with insurance companies that are expected to reimburse to hospitals the costs of medical care and services provided to insured patients as deductions. In many cases, insurance companies do not completely reimburse the expenses despite their contractual obligations. The total costs the companies evaded to pay amounted to about 10 percent of hospital revenues in 2000. Consequently, hospitals sometimes have to make up for their budget deficits by increasing the portion of the costs covered by the patient due to losses incurred by insurance companies [3].

In this study, the term 'health insurance deduction' is used to refer to the money not reimbursed by insurance companies for medical services provided by hospitals despite the contractual arrangements. Health insurance deductions happen mostly as the result of:

- Lack of proper documentation on the services provided by hospitals;
- Failure on the part of hospitals to submit full documents;
- Mismatch of the diagnostic-related group (DRG) system to calculate the true costs; and
- Additional services provided by hospitals such as drugs out of obligation, surgical services, unrelated diagnoses by doctors, and unrelated clinical tests.

Although, deductions could originate from different sources and for different reasons, this paper only focuses on hospital services and insurance obligations. For instance, insurance companies are obliged to reimburse the costs of delivery. In practice, if a mother is required to be hospitalized for more than 5 days, the costs for the extra days are not covered by the insurance companies. Or as another example, in the appendix surgery, insurance companies generally reimburse a certain amount of the cost that excludes the charges exacted under 'difficulty of surgery' [3].

The objective of this paper is to develop a DSS with a methodologically comprehensive and easy-to-use framework for the financial management of hospital to handle the health insurance deduction problem. The proposed framework is then validated through a case study of 200,000 insurance deduction documents over the period 2009-2010 from 150 different hospitals in Iran.

The rest of the paper is organized as follows: the following section provides a brief review of the literature. Section 3 briefly describes the decision support system, data mining, and multiple criteria decision making methods used as the main

methods along with the decision tree and TOPSIS methods employed in the case study. Section 4 presents the integrated framework proposed in this study. Section 5 describes a specific application of the proposed framework. Finally, the paper concludes with results and suggestions in Section 6.

II. Review of the Litreture

There are a variety of systems that can potentially support clinical decisions. Even Medline and similar healthcare literature databases can support clinical decisions. Decision support systems (DSS) have of long been incorporated into the healthcare information systems, but they usually have supported retrospective analyses of financial and administrative data [4, 5].

Basole et al. [6] developed a health advisor system which is a web-based game using organizational simulation in which players are tasked to manage people through the healthcare system by using various information, costs, and quality of care trade-offs with scores based on health outcomes and costs incurred. Gillies et al. [7] determined items that different stakeholder groups view to be important for inclusion in a DSS for clinical trial participation; with a view to use these as a framework for developing decision support tools in this context. North et al. [8] studied the research efforts in clinical DSS to compare triage documentation quality. Martínez-Pérezet al. [9] analyzed a sample of applications in order to draw conclusions and put forth recommendations about the mobile clinical DSS. Mobile clinical DSS applications and their inclusion in clinical practices have risen over the last few years. The authors found that the interface or its ease of use would impoverish the experience of the users if developers did not design them carefully enough.

Data Mining (DM) has been the most important tool used since 1990 for knowledge discovery from large databases. Recently, sophisticated DM approaches have been proposed for similar retrospective analyses of both administrative and clinical data [10, 11]. The use of DM to facilitate decision support provides a new approach to problem solving by discovering patterns and relationships hidden in the data, giving rise to an inductive approach to DSS. Roumani et al. [12] compared the performance of several common DM methods, logistic regression, discriminant analysis, Classification and Regression Tree (CART) models, C5, and Support Vector Machines (SVM) in predicting the discharge status of patients from an Intensive Care Unit (ICU). The non-expert users who tried the system obtained useful information about the treatment of brain tumors. Zandi [13] developed a bi-level interactive DSS to identify DM-oriented Electronic Health Record (EHR) architectures. The bi-level Interactive Simple Additive Weighting Model was then use to help medical decision makers gain a consensus on a DM-oriented EHR architecture. Bashir et al. [14] proposed the effectiveness of an ensemble classifier for computer-aided breast cancer diagnosis. A novel combination of five heterogeneous classifiers, namely Naïve Bayes, Decision tree using Gini Index, Decision Tree using information gain, Support Vector Machine, and Memory-based Learner were used to make the ensemble framework.

Remarkable progress has been made during the past 40 years in the Multiple Criteria Decision Making (MCDM) method so that it has nowadays developed into a mature discipline [15]. Recently, researchers have employed this method in a variety of areas including DM. Narci et al. [16] analyzed the effect of competition on technical efficiency through Data Envelopment Analysis (DEA) with five outputs and five inputs for the hospital industry in Turkey. Kusi-Sarpong et al. [17] introduced a comprehensive framework for green supply chain practices in the mining industry and presented a multiple criteria evaluation of green supply programs using a novel multiple criteria approach that integrates rough set theory elements and fuzzy Technique for Order Preference by Similarity to an Ideal Solution (TOPSIS). Aghdaei et al. [18] identified the synergies of DM and MADM and presented a wide range of interactions between these two fields from a new perspective. They provided an example of the integrated approach in supplier clustering and ranking.

Clearly, incorporation of DM and MCDM in decision support issues yields more powerful DSS since it offers more options for analysis, uses expert knowledge, and improves upon the process of analysis and evaluation [19].

III. Methodology

In this section, the method used in the proposed framework and its implementation such as decision support systems, data mining, decision trees, multiple criteria decision making, and TOPSIS are briefly described.

A. Decision Support System

Decision Support System (DSS) is a new computerized application serving organizational and business decision makers in their decision making process. The system is capable of extracting and collecting useful information from documents, business models, and raw data. It can even help solve problems and make useful decisions. The system is typically used for strategic and tactical decisions of a reasonably low frequency and high potential consequences for the upper-level management. The use of this system pays generously in the long run due to the short time taken for thinking through and modeling the problem [4, 5]. The three fundamental components of the DSS are as follows [20].

- A Database Management System (DBMS). DBMS serves as a data bank for DSS. It stores large quantities of data relevant to the class of problems for which the DSS has been designed and provides logical data structures through which the users interact.
- Model-base Management System (MBMS). The role of the MBMS is analogous to that of a DBMS. Its primary function is to keep specific models used in a DSS independent from the applications that use them.
- Dialog Generation and Management System (DGMS). The main product of an interaction of DGMS with a DSS is insight. As their users are often managers who are not computer-trained, DSS needs to be equipped with intuitive and easy-to-use interfaces.

B. Data Mining

Data mining (DM) is a popular technique for searching for and extracting interesting (i.e., non-trivial, implicit, previously unknown and unexpected potentially useful) and unusual patterns from data sources. DM problems are often solved by using a mosaic of different approaches drawn from computer science including multi-dimensional databases, machine learning, soft computing, and data visualization. Use is also made of statistics in terms of hypothesis testing, clustering, classification, and regression techniques [10, 11].

1) Decision Trees: A popular DM technique is the induction of decision trees. A decision tree (DT) is a machine learning technique used in classification, clustering, and prediction tasks. There are different tree-growing algorithms for generating DT such as C5.0, C&R trees, CHAID, and Quest[10, 11]. A DT starts from the root node which is one of the best attributes. Property values are then generated that correspond to each branch which generates a new node. For the best attributes according to the selection criteria, it uses an entropy-based definition of the information gain to select the test attribute within the node. The entropy characterizes the purity of a sample set. Suppose S is a set of data samples. We assume that the class label attribute has m different values, the definition of m different classes being C_i (i=1, ..., m), and set S_i is the number of samples in the class C_i . (1) is the sample classification based on expected information:

$$I(s_1, s_2, \dots, s_m) = -\sum_{i=1}^{m} p_i \log_2 \log_2(p_i) \quad (1)$$

where, P_i is the probability of any sample belonging to C_i, which is estimated using Si/S.

The set attribute A has ν different values $\{a_1, a_2, \dots, a_\upsilon\}$. A property can be divided into subsets $S\{s_1, s_2, \dots, s_\upsilon\}$, where Sj contains a number of S values in this sample and they have a value of aj in A. If we select the test attribute A, these subsets correspond to set S, which contains nodes derived from the growing branches. S_j assumes that S_{ij} is a subset of the samples of class C_i. Thus, A can be divided into subsets of entropy or expected information, which is given by (2):

$$E(A) = \sum_{j=1}^{\upsilon} \frac{s_{1j} + s_{2j} + \cdots + s_{mj}}{s} I(s_{1j}, s_{2j}, \dots, s_{mj}) \quad (2)$$

where, the item $(s_{1j} + s_{2j} + \cdots + s_{mj})/S$ subset is on the right of the first j and is equal to the number of subsets of the sample divided by the total number of S in the sample. (3) is a given subset for S_j:

$$I(s_{1j}, s_{2j}, \dots, s_{mj}) = -\sum_{i=1}^{m} p_{ij} \log_2(p_{ij}) \quad (3)$$

where, $P_{ij}=S_{ij}/|S_j|$ is a sample of S_j based on the probability of belonging to class C_i. (4) is a branch that will be used for encoding information.

$$Gain(A) = I(s_1, s_2, \dots, s_m) - E(A) \quad (4)$$

In other words, Gain(A) is attributable to a value of that property because of the expectations of the entropy of compression. Thus, a smaller entropy value leads to a lower correlation, whereas a higher corresponding information gain produces a subset of the division with a higher purity. Therefore, the test attribute DT selects the properties with the highest information gain. This creates a node and marks the property, where each value of the property creates a branch and divides the sample accordingly.

The DT contains leaves, which indicate the values of the classification variable, and decision nodes, which specify the test to be carried out. For each outcome of a test, a leaf or a decision node is assigned until all the branches end in the leaves of the tree [21, 22].

C. Multiple Criteria Decision Making

Multiple Criteria Decision Making (MCDM) is a sub-discipline of operations research that explicitly considers multiple criteria in decision-making environments. MCDM is concerned with structuring and solving decision and planning problems involving multiple criteria. In general, multiple criteria problems can be divided into two categories: Multiple Alternative Decision Making (MADM) and Multiple Objective Decision Making (MODM) problems. Typically, there is no unique optimal solution for such problems and it is necessary to use decision maker's preferences to differentiate between solutions [15, 23].

1) TOPSIS: The Technique for Order Preference by Similarity to an Ideal Solution (TOPSIS) method is a popular approach to MADM that has been widely used in the literature. Presented by Hwang and Yoon [23], it consists of the following steps [24, 25].

Step 1: The decision matrix is normalized through the application of (5):

$$r_{ij} = \frac{w_{ij}}{\sqrt{\sum_{j=1}^{J} w_{ij}^2}}, j = 1, 2, \dots, J; i = 1, 2, \dots, n \quad (5)$$

Step 2: A weighted normalized decision matrix is obtained by multiplying the normalized matrix by the weights of the criteria, (6):

$$V_{ij} = W_i^* r_{ij}, j = 1, 2, \dots, J; i = 1, 2, \dots, n \quad (6)$$

Step 3: PIS (maximum value) and NIS (minimum value) are determined by (7).

$$A^* = \{V_1^*, V_2^*, \dots, V_n^*\}, A^- = \{V_1^-, V_2^-, \dots, V_n^-\} \quad (7)$$

Step 4: The distance of each alternative from PIS and NIS is calculated using (8):

$$d_i^* = \sqrt{\sum_{j=1}^{J} (V_{ij} - V_j^*)^2},$$
$$d_i^- = \sqrt{\sum_{j=1}^{J} (V_{ij} - V_j^-)^2}, j = 1, 2, \dots, J \quad (8)$$

Step 5: The closeness coefficient for each alternative (CC_i) is calculated by applying (9):

$$CC_i = \frac{d_i^-}{d_i^* + d_i^-}, i = 1, 2, \dots, n \quad (9)$$

Step 6: At the end of the analysis, the ranking of alternatives is made possible by comparing CC_i values.

IV. THE PROPOSED FRAMEWORK

In this section, the proposed decision making framework for the health insurance deduction handling will be presented in detail.

To implement the integrated framework, an expert committee is first called in to extract a comprehensive list of healthcare services for patients in different cases, facilitated payments for the services, and an in-depth record of the actual treatments processed.

Fig. 1 shows the deployment diagram by integrating DSS, DM, and MCDM to make powerful, reliable, and efficient decisions in the insurance deduction handling. To facilitate the operations, the steps have been classified into four modules. Detailed descriptions of the modules and their steps are presented below.

A. Data Management Module

The hospital document system usually uses a computer system with a set of programs to track and store all the documents and instructions related to the health system [1-3]. These documents are usually provided by the hospital discharge, accounting, and statistical agencies and which should be considered as longitudinal registration data. The complete architecture of the data registration is shown in Fig. 2.

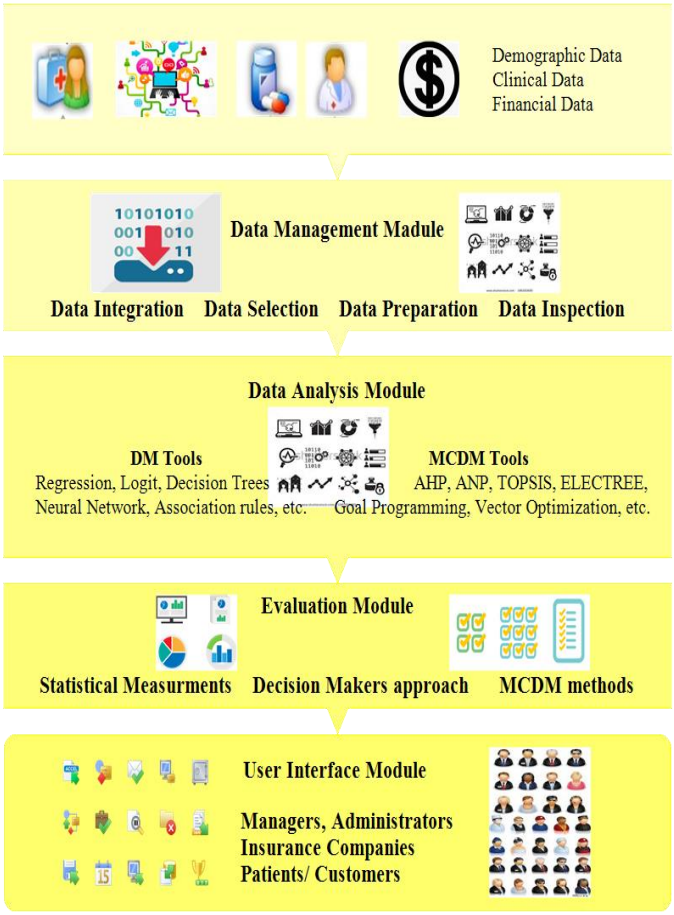


Fig. 1. The proposed framework

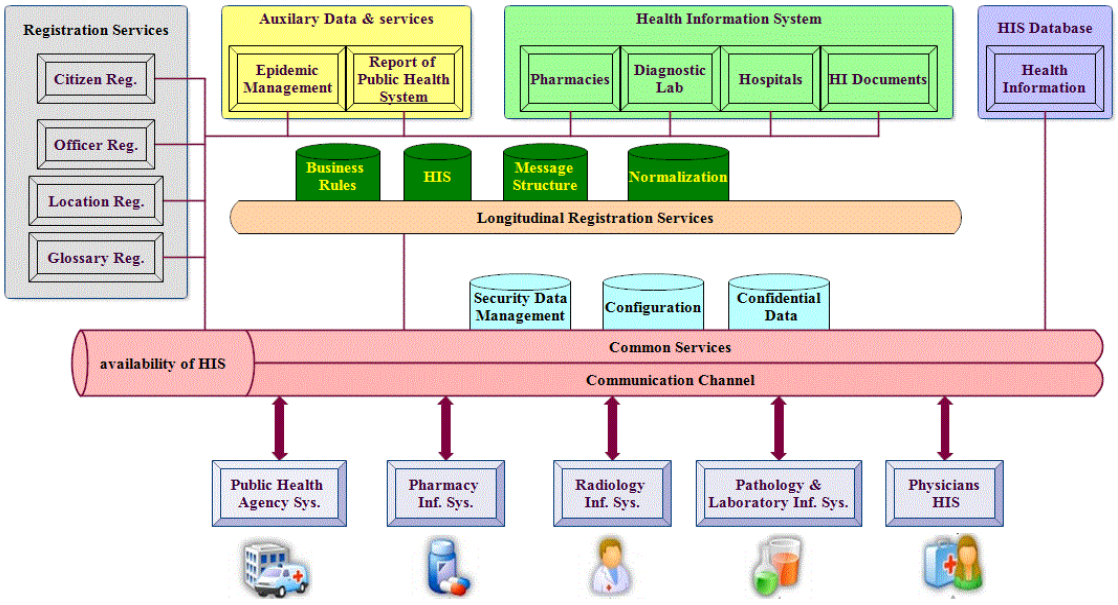


Fig. 2. The architecture of data registration in healthcare documents

In general, there are three categories of data for integration in hospital documents (Table 1):

- Demographic Data,
- Clinical Data, and
- Financial and Administrative Data.

TABLE I. SOME SELECTED ITEMS IN EACH CATEGORY OF HOSPITAL DOCUMENTS

Demographic data	Clinical Data	Financial and Administrative Data
• Gender	• Admitting doctor	• Bill services
• Birth date	• Medications	• Insurance
• Marital status	• Laboratory services	• Emergency treatment
• Education	• Surgery	• Discharge
• etc.	• etc.	• etc.

In the data selection step, not all the measured items should be selected from the database; unusable variables need to be discarded to save time and space while they may also yield wrong results which could be misleading to final users.

In the data preparation stage, the data are pre-processed and cleared for analysis. Examples of this are:

- Integrating the coding policy like DRG or ICD10-CM [26-29],
- Transforming some variables, such as the text data from the initial description of the pathology tests[30], and
- Dealing with missing and outlier data [30].

Data inspection is the final step in the data management module, in which the structure of the prepared data set is checked for the analysis of needs and their required tools.

B. Data Analysis Module

The objective of the data analysis module is to help hospital managers and insurance providers determine the characteristics of the relevant situations and predict future cases of insurance deductions by analyzing the available cases through a combination of DM and MCDM functions.

To overcome the existing problems, this module employs the data thus far prepared for:

- Classifying bill service data to predict actual treatment costs;
- Discriminating diseases to determine treatment costs;
- Using association rules and contingency tables of treatment costs and demographic data to study their possible relationships; and
- Using cluster analysis and frequent patterns to extract the patterns of causes of insurance deductions.

Moreover, financial and clinical analysis may be used:

- To study the relationships among the tests for specific diseases prescribed by different physicians using association rules and frequent pattern recognition. This leads to the identification of efficient from non-efficient tests, the results of which can be used for cost management and determination of the rate of unrelated diagnoses by each physician.
- To classify all types of services offered in order to identify the necessary orders and supplies such as drugs, visits by physicians and specialists, and pathology tests to support administrative functions;
- to evaluate the priority of development activities in hospitals based on prioritized utility functions; and
- To predict total hospital expenditures for different seasons and months using temporal mining and time series analysis.

C. Evaluation module

Depending on the type of analysis required, use will be typically made of statistical criteria, training and test datasets, cross validations, or the like for the evaluation of the results obtained.

Furthermore, the proposed framework uses MCDM techniques and decision maker opinions for evaluation. For instance, MADM methods such as AHP, ANP, ELECTREE, and TOPSIS could be employed to evaluate and rank the results. Programming and genetic algorithms will be more efficient when using a scoring system for performance and optimization as in the assessment of insurance deductions.

D. User interface module

The user interface module should present a comprehensive view of the decision making process depending on the requirements put forth by managers and administrators. Poor usability is one of the core barriers to adoption of a system, acting as a deterrent to DSS routine use.

Generally, the following points should be considered in the design of the interface for the hospital document management system regarding the insurance deductions handling:

- Monitoring the data collection process and its integration;
- Monitoring each step of the data management module;
- The possibility for employing different DM and MCDM methods for each type of data depending on the objectives of data analysis and inspection;
- Presenting the results in accordance with administrators' needs and requirements;
- The possibility for evaluation of the results obtained from the data analysis module including MCDM methods; and

- The possibility for sensitivity analysis and evaluation of several scenarios by decision makers.

V. IMPLEMENTATION

In this section, the efficiency of the proposed framework is investigated by using it to predict the most likely services which lead to insurance deductions in different hospitals. For this purpose, the information from 200,000 documents for patients hospitalized in 150 different hospitals over the period from 2009 to 2010 is integrated to create around 97,532 records.

In the data selection step, different types of hospitals were considered. Also, the documents were chosen using International Classification of Diseases (ICD), ignoring emergencies and accident cases.

In addition, transformation and normalization were used in the data preparation process. As most of the records included very low deductions, biases of the model were avoided by considering ROD as zero if the rate of deduction (ROD) was less than 3 percent. The selected features after data inspection are presented in Table 2.

TABLE II. SELECTED FEATURES

Feature	Range
Age	1–107 years old
Sex	female/ male
Hospitalization	1–212 days
ICD code	35 popular diagnostics
Hospital type	public, private, academic, charity
Service types	27
Rate of deduction	The ratio of deduction to the total amount

Table 3 presents the distribution of deductions according to types of services. As can be seen, almost half the insurance deductions belonged to medications, laboratory test charges, and supplies used.

TABLE III. DISTRIBUTION OF DEDUCTIONS ACCORDING TO SERVICE TYPES

Type of service	Frequency	Relative frequency	Total amount of deductions (Rials)
Medication	111879	22.8	9,425,106,171
Laboratory tests	60973	12.43	3,695,746,116
Supplies used	57640	11.75	5,132,008,336
Operating Room (OR)	20208	4.12	7,458,749,993
Supplies used for OR	38198	7.79	6,463,764,310
Surgery	26102	5.32	2,906,490,815
Physicians	25361	5.17	592,919,118
Nurses	22770	4.64	6,661,996,161
Bed	22508	4.59	3,779,407,806
Anesthetics	20841	4.25	3,188,967,335
Total	406480		49,305,156,161

The focus here was on the data analysis module. Given the goal of decision making, the Decision Tree (DT) was exploited to predict insurance deductions from types of hospital services [10, 11]. In this case, the algorithms of C5.0, C&R trees, CHAID, and Quest were applied and a 10-fold cross validation was used. Also, for estimating the performance of the predictive models, the records of 2009 (about 53,795 cases) were used as the training dataset while those of 2010 were used as the validation dataset. The results obtained are reported in Table 4.

TABLE IV. COMPARISON OF THE RESULTS OBTAINED FROM THE ALGORITHMS USED

Algorithm	Overall accuracy	Accuracy related to deduction class	Accuracy related to no deduction class	Number of extracted rules
C5.0	86.72	88.05	86.43	1640
C&R trees	86.78	89.84	82.43	62
CHAID	83.10	89.79	81.62	6
Quest	86.12	83.53	86.69	13

As the purpose of this analysis was to extract reliable, useful, and meaningful rules for managers and administrators of hospitals and insurance providers, the huge number of patterns (1721 rules) discovered did not seem sensible or usable. The human brain is reportedly incapable of processing a large number of logical phrases and rules as it will be hard for it make good sense out of it [31, 32]. The evaluation step was, therefore, applied to prioritize the rules extracted. In this study, certain important performance measures were initially defined and the TOPSIS method was used to rank the rules that could be extracted. Thus, the following concepts were defined as performance measures:

- Accuracy (ACC): The correct classification rate of the rule based on the test dataset.
- Stability (STAB): Not a great variation is allowed in the accuracy rate when a rule is applied to different datasets. Thus, one might minimally expect that a rule does not exhibit a great variation when applied for the validation dataset or the training dataset. Then, STAB = Min { ACC_t / ACC_v , ACC_v / ACC_t }.
- Simplicity (SIMP): This limits the number of attributes in a rule.
- Discriminatory Power (DP): The ratio of discriminated cases for the rule; ideally one would like to have rules (leaves) that are totally pure (i.e., all the classes except for one has a zero probability for each leaf) but in many cases this does not occur and so the class that is associated with the rule (leaf) is simply the class with the largest frequency for the given rule based on the training dataset.
- ROD: The ratio of deduction of the rule to the total amount.

As already mentioned, the best alternative in the TOPSIS approach is the one nearest to the ideal solution and the one farthest from the negative ideal solution. Also, it is assumed

that all the criteria have identical weights and importance. Table 5 presents brief calculation results of this method.

TABLE V. CALCULATION OF THE TOPSIS METHOD

	ACC	STAB	SIMP	DP	ROD	d*	d^-	CC
Rule # 1	0.29	0.33	0.27	0.16	0.05	1.62	0.87	0.35
Rule # 12	0.31	0.61	0.57	1	0.58	1.06	1.41	0.57
Rule # 123	0.81	0.73	0.43	0.66	0.41	0.87	1.46	0.63
Rule # 1234	0.19	0.54	0.68	0.80	0.03	1.52	1.03	0.41

In this Table, the columns for the criteria defined are normalized scores of each rule, d* is the deviation from the ideal alternative, d^- is the deviation from the negative ideal alternative, and CC is the relative closeness to the ideal solution. All the rules were then sorted based on the CC column from the TOPSIS calculation and the most important rules were extracted for planning and decision making by managers and administrators of hospitals and insurance providers. Some of the results are presented in Table 6.

TABLE VI. THE FINAL RESULTS

CC	Cases	Record of deduction
0.94	%35	Supplies used for a patient with heart disease and overnight hospitalization
0.85	%30	OR's supplies used for a patient with cataract and overnight hospitalization
0.73	%30	Bed for a labour patient a 7-day hospitalization period

Using these rules and information, hospital managers can revise their policies for similar cases as to how to reimburse the expenses of their medical services and to negotiate with insurance providers on how to deal with insured patients receiving similar services. Moreover, insured patients can in this way be fully informed about the services covered by insurance companies.

VI. CONCLUSIONS

Hospital management is a most complex decision making process that has to deal with huge arrangements related to such financial and administrative process, medical operations, and the patient services, etc. The decision support system is an effective technology that makes it possible to properly respond to such hospital management requirements.

One major challenge commonly arising between insurance companies and hospital managers is the disputes and disagreements over the reimbursement of medical expenses of insured patients. A majority of hospital revenues are reclaimed as per contracts with insurance companies which provide insurance policies.

The 'health insurance deduction' is referred to the money not reimbursed by insurance companies for medical services provided by hospitals despite the contractual arrangements.

This paper presented an integrated framework for handling health insurance deduction based on DSS, DM, and MCDM methodologies.

Nowadays, decision makers invariably need to use DSS to tackle complex decision making problems. In this area, DM plays an important role in extracting valuable information. Also, MCDM method deals with such varied areas as choosing the best option among various alternatives and optimizing goals among multi-objective situations.

The proposed framework is capable of achieving enhanced decision making performance, improving the effectiveness of solutions developed, and enhanced possibilities for tackling new types of problems not addressed before. Application of the proposed method to a case study yielded objective and comprehensive results which assist hospital managers to negotiate with insurance providers on how to handle the insurance deduction.

In the forthcoming work, we will apply the proposed framework in other aspect of hospital management, medical diagnosis and possibly other applications in the near future.

REFERENCES

[1] W.J. Hopp, WS. Lovejoy, "Hospital Operations: Principles of High Efficiency Health Care", First edition, Pearson FT Press; 2012.

[2] J.R., G. Langabeer, "Health Care Operations Management: A Quantitative Approach To Business And", First Edition, Jones & Bartlett Learning; 2007.

[3] http://www.centinsur.ir.

[4] R A. Greenes, "Clinical Decision Support, The Road to Broad Adoption", Second Edition, Academic Press, 2014, doi:10.1016/B978-0-12-398476-0.00032-4.

[5] C. Jao, "Decision Support Systems", InTech, 2012, DOI: 10.5772/3371.

[6] E. S. Berner, "Clinical Decision Support Systems, Theory and Practice", Second Edition, Springer, 2007.

[7] R.C. Basole, D.A. Bodner, W.B. Rouse, Healthcare Management through Organizational Simulation: An Approach for Studying Emerging Organizational Ideas and Concepts, *Decision Support Systems*, 55 (2): 552-563, 2013.

[8] K. Gillies, Z. Skea, S. MacLennan, C. Ramsay, M. Campbell, Determining items for inclusion in a decision support intervention for clinical trial participation: a modified Delphi approach, *Trials*, 2013, DOI: 10.1186/1745-6215-14-S1-O64.

[9] F. North, D.D. Richards, K.A. Bremseth, Clinical decision support improves quality of telephone triage documentation - an analysis of triage documentation before and after computerized clinical decision support, *BMC Medical Informatics and Decision Making*, 2014, DOI:10.1186/1472-6947-14-20.

[10] B. Martínez-Pérez, I. Torre-Díez, M. López-Coronado et al., Mobile Clinical Decision Support Systems and Applications: A Literature and Commercial Review, *J of Medical Systems*, 38:4, 2014, DOI:10.1007/s10916-013-0004-y.

[11] M. Kuntardzic, "Data Mining: Concepts, Models, Methods and Algorithms", 2nd Edition, Wiley, 2011.

[12] J. Han, M. Kamber, J. Pei, "Data Mining", 3rd Edition, Elsevier Pub, 2012.

[13] Y.F. Roumani, J.H. May, D.P. Strum, L.G. Vargas, Classifying highly imbalanced ICU data, *Health Care Management Science*, 16 (2) 119-128, 2013.

[14] F. Zandi, A bi-level interactive decision support framework to identify data mining-oriented electronic health record architectures, *Applied Soft Computing*, 18: 136–145, 2014.

[15] S. Bashir, U. Qamar, F. Hassan Khan, Heterogeneous classifiers fusion for dynamic breast cancer diagnosis using weighted vote based ensemble, *Quality & Quantity*, 2014, DOI: 10.1007/s11135-014-0090-z.

[16] A. Ishizaka, P. Nemery, "Multi-Criteria Decision Analysis: Methods and Software", Wiley, 2013.

[17] H.Ö. Narcı, Y.A. Ozcan, İ. Şahin, et al., An examination of competition and efficiency for hospital industry in Turkey, *Health Care Management Science*, 2014, DOI: 10.1007/s10729-014-9315-x.

[18] S. Kusi-Sarpong, C. Bai, J. Sarkis, X. Wang, Green supply chain practices evaluation in the mining industry using a joint rough sets and fuzzy TOPSIS methodology, *Resources Policy*, 2014, DOI:10.1016/j.resourpol. 2014.10.011.

[19] M. H. Aghdaiea, S. H. Zolfanic, E. K. Zavadskas, Synergies of Data Mining and Multiple Attribute Decision Making, *Procedia - Social and Behavioral Sciences*, 110 (24): 767–776, 2014.

[20] S. Khademolqorani, A.Z. Hamadani, An Adjusted Decision Support System through Data Mining and Multiple Criteria Decision Making, *Procedia - Social and Behavioral Sciences*, 73, 388 – 395, 2013.

[21] S. Chaudhuri, U. Dayal, V. Ganti, Decision support system components, *Computer, IEEE*, 34 (12): 48–55, 2001.

[22] J.R. Quinlan, R.L. Rivest, Inferring decision trees using the minimum description length principle, *Informat Computat*, 80 (3) 227–248, 1989.

[23] L. Breiman, J.H. Friedman, R.A. Olshen, C. J. Stone, "Classification and regression trees", Belmont: Wadsworth. Statistics probability series, 1984.

[24] C.L. Hwang, K. Yoon, "Multiple Attribute Decision Making Methods and Applications", Springer, 1981.

[25] E. Afful-Dadzie, S. Nabareseh, A. Afful-Dadzie,. Z. K. Oplatková, A fuzzy TOPSIS framework for selecting fragile states for support facility, *Quality & Quantity*, 2014, DOI: 10.1007/s11135-014-0062-3.

[26] T. Wachowicz, P. Błaszczyk, TOPSIS Based Approach to Scoring Negotiating Offers in Negotiation Support Systems, *Group Decision and Negotiation*, 22 (6) 1021-1050, 2013.

[27] O. Ben-Assuli, I. Shabtai, M. Leshno, S. Hill, EHR in Emergency Rooms: Exploring the Effect of Key Information Components on Main Complaints, Journal of Medical Systems, 37: 9914-9936, 2014, DOI: 10.1007/s10916-014-0036-y.

[28] D. Fraser, B.A. Christiansen, R. Adsit et al., Electronic health records as a tool for recruitment of participants' clinical effectiveness research: lessons learned from tobacco cessation, *Translational Behavioral Medicine*, 3 (3): 244-252, 2013.

[29] M. Taboada, M. Meizoso, D. Martínez et al., Combining open-source natural language processing tools to parse clinical practice guidelines, *Expert Systems*, 30 (1): 3–11, 2013.

[30] J. Kline, "Regulatory Requirements and Health Care Codes", Handbook of Biomedical Engineering, 1988.

[31] L. Page, "The transformation of the hospital call center", COR Healthcare Market Strategist, 2004.

[32] B. Whitworth, Some Implications of Comparing Brain and Computer Processing, *Proceedings of the 41st Hawaii Int Conf on Sys Sciences, IEEE*, 2008.

[33] R. Marois, J. Ivanoff, Capacity limits of information processing in the brain, *TRENDS in Cognitive Sciences*, 9 (6): 296-305, 2005.

A Multi-Label Classification Approach Based on Correlations Among Labels

Raed Alazaidah
Computer Science Department
Philadelphia University
Amman- Jordan

Fadi Thabtah
E-Business Department
Canadian University of Dubai
Dubai-UAE

Qasem Al-Radaideh
CIS Department
Yarmouk University
Irbid- Jordan

***Abstract*—Multi label classification is concerned with learning from a set of instances that are associated with a set of labels, that is, an instance could be associated with multiple labels at the same time. This task occurs frequently in application areas like text categorization, multimedia classification, bioinformatics, protein function classification and semantic scene classification. Current multi-label classification methods could be divided into two categories. The first is called problem transformation methods, which transform multi-label classification problem into single label classification problem, and then apply any single label classifier to solve the problem. The second category is called algorithm adaptation methods, which adapt an existing single label classification algorithm to handle multi-label data. In this paper, we propose a multi-label classification approach based on correlations among labels that use both problem transformation methods and algorithm adaptation methods. The approach begins with transforming multi-label dataset into a single label dataset using least frequent label criteria, and then applies the PART algorithm on the transformed dataset. The output of the approach is multi-labels rules. The approach also tries to get benefit from positive correlations among labels using predictive Apriori algorithm. The proposed approach has been evaluated using two multi-label datasets named (Emotions and Yeast) and three evaluation measures (Accuracy, Hamming Loss, and Harmonic Mean). The experiments showed that the proposed approach has a fair accuracy in comparison to other related methods.**

Keywords—Classification; Data mining; Multi-label Classification

I. INTRODUCTION

Data classification is a form of data analysis that can be used to extract models describing important data classes. The classification task concentrates on predicting the value of the decision class for an object among a predefined set of classes given the values of some given attributes for the object. In general, data classification is a two-step process. In the first step (learning), a model that describes a predetermined set of classes or concepts is built by analyzing a set of training database objects. Each object is assumed to belong to a predefined class. In the second step, the model is tested using a different data set.

Classification problems can be divided into three main categories: Binary classification, Multi-Class classification and Multi-Label classification. In binary classification, there are only two possible values for the class label (X, Y). However, most real world application domains contain several classes and therefore several multi-class approaches have been proposed.

Formally, the traditional classification problem can be defined as follows: "let D denotes the domain of possible training instances, and Y be a list of class labels, let H: $D \rightarrow Y$ denotes the set of classifiers. Each instance $d \in D$ is assigned a single class label y that belongs to Y. The goal is to find a classifier $h \in H$ that maximize the probability that $h(d) = y$, for each test case (d , y). In multi- label problem , however , each instance $d \in D$ can be assigned multiple labels $y_1, y_2, \dots , y_k$ *for* $y_i \subseteq Y$, and is represented as a pair (d , ($y_1, y_2 ,\dots ,y_k$)) where ($y_1, y_2, \dots ,y_k$) is a list of ranked class labels from Y associated with the instance d in the training data [1].

Multi label classification is concerned with learning from set of instances that are associated with a set of labels, that is, an instance could be associated with multiple labels at the same time. This task occurs frequently in application areas like text categorization, multimedia classification, bioinformatics, protein function classification and semantic scene classification. An Example of a multi label dataset is presented in Table1. In practice, most of the current classification approaches do not consider the generation of rules with multiple labels from multi-class or multiple label data [2].

TABLE I. MULTI-LABEL DATA

A1	A2	A3	A4	Class
5	A	2	R	X, Y
3	B	0	A	X, W, Z
3	B	2	A	Z
3	B	6	T	Y, Z

This paper proposes a guided multi-label classification approach based on correlations among labels in class label attribute and then applying a classical classification algorithm to learn rules from the training dataset. Most of multi-label classifications methods, both problem transformation methods and algorithm adaptation methods depend, for its classification task, on a function that maps between the attributes and the labels in the training data. The proposed approach introduces a new approach to solve the problem of multi-label classification. This approach is based on correlations among labels learned by predictive classification.

II. Related Work: Multi-Label Classification Methods

Existing methods for handling multi-label classification can be grouped into two main groups. The first group, which is an algorithm independent, is called problem transformation methods, while the second group is an algorithm dependent, and is called algorithm adaptation methods. The first group transforms multi-label classification problem into one or more single classification problem, while the second group extends a specific learning algorithm, in order to handle multi-label data directly [3].

A. Problem Transformation Methods

Several problem transformation methods exist in the literature that is used to convert multi-label classification problem into one or more single label classification problem. To exemplify these methods, we will use the dataset of Table2 which consists of four examples that belong to the following class set {Reading, Swimming, Painting, TV Watching}

TABLE II. Multi-Label Data Set

Inst #	Reading	Swimming	Painting	TV Watching
1		X	X	
2	X		X	
3		X		X
4		X		

The first problem transformation method discards every multi-label instance from the data set. Therefore, in the previous example, instances 1, 2, 3 will be discarded. Another problem transformation method selects one of the multiple-labels of each multi-label instance either randomly or subjectively. The transformed version of the previous example instances is presented in Table3.

TABLE III. Multi-Label Data Set

Inst #	Reading	Swimming	Painting	TV Watching
1		X		
2			X	
3				X
4	X			

The copy transformation method transforms every multi-label instance to a single label instance by replacing the multi-label instance (xi, yi) with $|y_i|$ instances. Several transformation methods could be then chosen such as: (1) copy-weight which associates a weight of $(1/|y_i|)$ to each of the transformed examples, (2) select-max (most frequent), (3) select-min (least frequent), (4) select-random, and (5) the ignore transformation option.

One of the most popular transformation methods, that learn single binary classifier for every label in the label set, is called Binary Relevance (BR) [3]. This method transforms the original data set into |L| data sets, which contain all the instances from the original data set. It then gives a positive sign for a label, if it exists in the data set and a negative sign otherwise. To classify new instance, the BR method returns the union of all labels that are predicted by the $|L|$ classifiers.

Although Binary Relevance is a simple transformation method, it is based on implicit assumption of labels independence which might be completely incorrect in the data.

Another method called the Label Power Set (LP) is a straight forward method that works as follows: it considers each unique set of labels that exists in the data set as a new single label in single – label classification task as shown in Table4.

TABLE IV. Multi-Label Data Set

Inst #	Label
1	{Swimming, Painting}
2	{Reading, Painting}
3	{Swimming , TV Watching}
4	{Swimming}

To predict the class label of a new instance, the LP method returns the most probable class which actually could be a set of labels in the original data set [4]. The Computational complexity of LP is upper-bounded by (min ($|L|$, 2^k)) where k: is the total number of classes in the data set before transmission, and usually it is much less than 2^k. LP has an advantage of taking labels correlations into account, on the contrary of BR, but it has a disadvantage when a large number of classes in the original data set associated with small number of instances, which may cause an imbalance problem for learning.

The previous mentioned problem of LP was addressed by the pruned problem transformation methods [5] which used a user- defined threshold to prune some label sets that occur less than this threshold. The pruned set could be replaced by disjoint subsets of these labels that are more frequent in the data set.

The RAKEL (Randon K label sets) method is an effective transformation method that breaks the initial set of labels into a number of small random subsets called label-sets and then employs the LP method to train a corresponding classifier, where k is a parameter that determines the size of the subsets [4]. RAKEL offers advantages over LP for the two reasons: (a) The resulting single label classification tasks are computationally simpler, and (b) The resulting single label classification tasks are characterized by much more balance distribution of class values. In RAKEL, the parameter K which is used to determine the size of the subsets and specified by the user should be small to avoid the problems associated with the LP method.

The Ranking by Pair wise Comparison (RPC) approach by [6] transforms the multi-label classification problem into a single label classification problem through performing pair wise comparisons of labels. RPC learns ($|L|$ * ($|L|$ - 1)) / 2 binary classifiers, one model for each different pair of labels. For predicting new instance, all models are invoked and ranking is obtained through counting the votes received by each label. An extension of RPC called Calibrated Label Ranking (CLR) [7] which introduces a virtual label (often called calibration label, L_0) that aims to separate relevant labels from irrelevant ones.

Another problem transformation method called the Classifier Chains (CC) method tries to enhance the BR method through taking label correlations into account [8]. CC builds |L| binary classifier for each label as in BR. Then Classifiers are linked along a chain where each classifier deals with the binary relevance problem associated with label $l_j \in L$. The feature space of each line in the chain is extended with 0/1 label association of all previous links. The CC method counteracts the disadvantages of the binary method while maintaining acceptable computational complexity.

The Ensemble of Classifier Chains (ECC) method is an enhancement version of CC which in turn is an enhancement of BR. ECC trains *m* Classifier Chains $C_1, C_2, \dots, C_m$, Where each C_k is trained with a random chain ordering of L and a random subset of D. Each C_k model is likely to be unique and able to give different multi label predictions. These predictions are then summed by label so that each label receives a number of votes. A threshold is used to select the most popular labels which form the final prediction of multi label set [8].

Another problem transformation method called Pruned Sets (PS) is an enhancement of Label Power-set (LP) which treats every unique subset of labels as a single label, and suffers from label imbalance specially, when number of training examples is small and number of labels is too large [5]. PS try to solve this problem by focusing only on the most important correlations, which reduce complexity and improve accuracy [8].

B. Algorithm Adaptation methods

Algorithm Adaptation methods extend a specific single label learning algorithm in order to handle multi-label data directly. In this section, we introduce a brief plethora of algorithm adaptation methods grouped by the learning concept that they extend.

Reference [9] developed a re-sampling technique and modified the C4.5 algorithm to deal with a gene hierarchy multi-label classification problem.

Reference [1] proposed a Multi-class, Multi-label Associative Classification algorithm (MMAC) which is an associative rule learning based covering algorithm that recursively learns a new rule and each time removes the examples associated with that rule. Labels for the test instances are ranked according to confidence, support, and rule's cardinality (number of conditions in the left hand side (LHS) of the rule).

Reference [4] proposed the AdaBoost.MH and AdaBoost.MR as two extensions of AdaBoost for multi-label data, where AdaBoost.MH aims to reduce Hamming loss and AdaBoost.MR aims to increase accuracy.

Reference [10] proposed a *K*- nearest Neighbors (*K*NN) lazy learning based method for multi label data. In general, the KNN based methods share the same first step with *K*NN (retrieving the *K* nearest example) and differ from each others on the aggregation of the label sets of these examples.

III. THE PROPOSED APPROACH FOR MULTI LABEL CLASSIFICATION

The general structure of the proposed approach consists of three phases: (a) Transforming multi-label dataset into single label dataset and discovering correlations among labels. (b) Applying a rule-based classification algorithm on the transformed dataset. (c) Generating the multi-label rules based on the output of the rule-based classifier and the correlations among labels. Fig.1 shows the general structure of the proposed approach and the main steps of the approach are described in Fig. 2.

As shown in Fig. 1, the input of the algorithm is a multi-label dataset, and then two operations are performed on the multi-label dataset: the first operation is transforming multi-label dataset into a single label dataset; in this step there are several methods to choose from such as: selecting the most frequent label, selecting the least frequent label or select any label randomly.

For the proposed approach we choose to select the least frequent label as transformation criteria. The second operation is to find all positive association among labels using the predictive Apriori method [11]. This operation tries to associate each label with labels from the label set; if that is possible. The output after performing these two operations will be:

1) A single label dataset which has been extracted or transformed from multi-label dataset using the least frequent label criteria.

2) Rules between labels with different rule's cardinality, starting from cardinality 1 up to rule's cardinality which is equal to the dataset cardinality -1, (i.e, Association rule's cardinality = Label Cardinality – 1).

In the next step, a single rule-based classifier is applied on the transformed dataset. Several rule-based classifiers could be used in this stage such as RIPPER, IREP, PART or Prism. The output of any single rule based classifier will be set of "*IF-THEN*" rules with one consequent on the right-hand-side of the rule like the following rule:

$$IF\ (con_1\ and\ con_2\ and\ \dots\ con_n)\ THEN\ Label.$$

Using both, the output of the single rule based classifier and the rules based on the correlations among labels previously discovered, we will be able to build a multi-label rules classifier in the form:

$$IF\ (con_1\ and\ con_2\ and\ \dots\ con_n)\ THEN\ \{Label_1, Label_2, \dots, Label_n\}.$$

The Learning Phase

The learning Phase in the proposed approach consists of two different tasks. The first task is an unsupervised learning task, which aims to discover the correlations among labels using Predictive Apriori. While the second task is a supervised learning task that aims to predict the class label of unseen instance as accurate as possible using a rule based classifier.

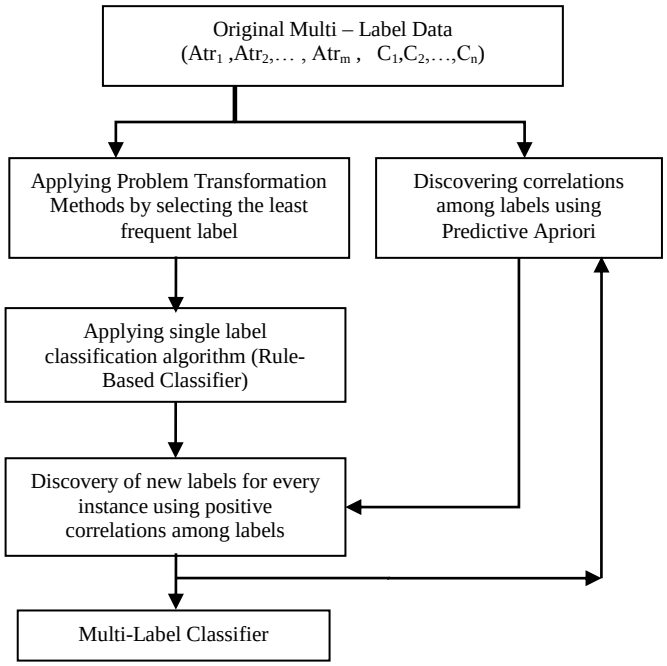


Fig. 1. The General structure of the proposed approach

Algorithm 1: Multi Label Classification Approach based on Correlation among Labels (MLC-ACL)

Input: Multi-label dataset as training data.
Output: A set of Multi-Label rules.
Phases:
Phase 1: Dataset Transformation
a. Transforming multi-label dataset into a single label dataset by selecting the least frequent label associated with each training instance.
Phase 2: Learning
a. For every label in the label set of the dataset, find the highest accuracy positive rule in the form of: IF label X exists THEN label Y exists.
b. Applying a rule based classifier on the transformed data set and producing the rules set.
Phase 3: Classification
a. Generating the multi-label rules set, using the single rules set produced by the classifier in Phase 2, and the associative rules for each instance that has been discovered in phase 1.
b. Use the multi-label rules set (Classifier) for Prediction.

Fig. 2. The main phases of the Proposed Algorithm

A. Discovering of Positive Correlations among Labels.

Suppose we have the itemsets (Labels) C1, C2, and C3. We are interested in having association rules with good confidence between every possible Pair-wise of the three previous labels. For the first two labels C1, C2 we may have the following rules for example:

1- If C2=1 Then C1=0
2- If C1=1 Then C2=1

In the proposed approach, we are interested in rules like the second rule, we are looking for a rule in a form of (*IF Label x exists THEN label y exists*). For each label (*x*) in the dataset we want to find another label (*y*) that has a positive correlation with it, i.e. label (*x*). In case we have more than one label positively associated with the label in the antecedent, we select the rule with the highest confidence or accuracy. For example suppose that we have the following association between *C1*, *C2* and *C3*:

1- If C1 =1 Then C2=1 (Accuracy = 0.80)
2- If C1=1 Then C3=1 (Accuracy = 0. 71)

In the previous case, we choose the rule with the highest accuracy, so rule one will be selected, and rule two is ignored. In fact ignoring such a rule with a meaningful confidence such 0.71 may cause too much information loss but let us stuck on the choice of selecting the best rule, and leave ignoring other rules with meaningful confidence to be discussed later in the future work section.

After having all positive associations of length "1" between labels in the dataset , we move forward to find all positive associations of cardinality "2" as the following rule (If C1=1 and C2=1 Then C3=1) and so forth.

For the proposed approach, we will choose the rule with the highest accuracy without any pre specified condition about the value of accuracy, such as the accuracy should be greater than or equals to a predefined user threshold. For example, suppose we have the following rules:

1- If C1=1 Then C2=1 (Accuracy = 0.27)
2- If C1=1 Then C3=1 (Accuracy = 0.19)

B. Applying Rule-Based Classifier

After having the transformed data set, and finding the highest positive association rules among labels, we are ready to apply any single rule-based classification algorithm to the transformed data, and we choose PART classifier.

PART is a rule-based classification algorithm that combines between two approaches. The first one is creating rules using decision tree, and the second one is separate and conquer learning method [12]. The algorithm produces accurate rules in the same size as those generated by decision tree C4.5 algorithm. PART algorithm has been chosen for being accurate, efficient and fast.

The Prediction Phase

Finally, the classifier consists of a Set of multi-label rules that have been learned from both correlations among labels and rule-based classifier. This classifier will be used in the prediction step to predict the class label / labels of a new instant.

IV. AN ILLUSTRATIVE EXAMPLE FOR THE PROPOSED APPROACH

For more clarification, this section presents a complete step by step example for the proposed approach using the "Emotions" dataset which has been downloaded from the following address (http://mulan.sourceforge.net/datasets.html). The characteristics of the dataset are presented in Table5. Table6 shows the frequency of the six labels in the “emotions” dataset. It is clear that the Most Frequent Label (MFL) is "Relaxing" and the Least Frequent Label (LFL) is "Quite-still".

TABLE V. MULTI-LABEL DATASET INFORMATION

Dataset name	Domain	# of Instances	# of Numeric Attributes
Emotions	Music	593	72

TABLE VI. "EMOTIONS" DATASET LABELS STATISTICS

Label	Amazed	Happy	Relaxing	Quite-still	Sad	Angry
Frequency	173	166	264	148	168	189

A. Approach Phases

Here we describe the main phases of the approach as the following:

Phase1 (a): Transform the dataset ("Emotions") into a single label dataset using least frequent label. Sample of the transformed dataset is presented in Table7. As we can see in Table7, the first example is associated with three labels at the same time (*Relaxing*, *Quite-Still*, *Sad*), and since "*Quite-Still*" has frequent 148 which is less than the frequent of "*Relaxing*" (264) and "Sad" (168), it will be transformed to the single label "*Quite-Still*". The second example is associated with two labels: "*Amazed*" with frequent equals to 173 and "*Angry*" with frequent 189, so it was transformed to the least frequent label which is "*Amazed*", and so on for the rest of examples.

TABLE VII. TRANSFORMING "EMOTIONS" DATASET INTO SINGLE LABEL DATASET

Ins #	Amazed	Happy	Relaxing	Quite-still	Sad	Angry	Class
1	0	0	1	1	1	0	**Quite-still**
2	1	0	0	0	0	1	**Amazed**
3	0	0	0	0	1	0	**Sad**
4	0	1	1	0	0	0	**Happy**
5	0	0	0	0	1	0	**Sad**
6	0	0	1	0	1	0	**Sad**

Phase1(b): The second step is to find positive correlations among labels using predictive Apriori. Best correlations are chosen without determining any threshold value in this stage, and since "Emotions" dataset is of cardinality "2"; association rules will be with "1" condition only in the antecedent. Table11 shows the complete positive correlations among labels in "Emotions" dataset.

As notices in Table8, Rule #5 has the lowest accuracy, in this case we will stuck in the choice of having the highest positive association among labels, and since no other rule could be found to be associated with the label "angry", and has accuracy greater than this rule, this rule is chosen.

TABLE VIII. POSITIVE CORRELATIONS AMONG LABELS IN "EMOTIONS" DATASET

Rule #	Rule	Accuracy
1	*IF amazed THEN angry*	0.53
2	*IF happy THEN relaxing*	0.44
3	*IF Quite-still THEN sad*	0.71
4	*IF Sad THEN Relaxing*	0.57
5	*IF angry THEN Relaxing*	0.03
6	*IF Relaxing THEN Relaxing*	1.00

Phase (2): The third step in the proposed approach is to apply a rule based classification algorithm on the transformed dataset. Table9 shows some of the learning rules discovered after applying the PART classifier.

TABLE IX. LEARNING RULES DISCOVERED AFTER APPLYING THE PART CLASSIFIER

Rule #	Rule Condition	Consequence
1	IF AQ > 0.217678 AND B <= 0.090652 AND V > 0.580398 AND AZ > 3.787686 AND AX > 0.060033 AND BD <= 0.173826	**Sad**
2	IF AQ <= 0.215792 AND BJ <= 0.108461 AND J <= 1.021892 AND BO <= 0.066288	**Angry**
3	IF AS > 0.206592 AND AI > 0.010202 AND D > -76.700621	**Amazed**
4	IF AS > 0.206592 AND AI > 0.010202 AND B <= 0.191563	**Quit-Still**
5	IF AS > 0.208738 AND B <= 0.119991 AND AP > 0.213677 AND BN <= 102 AND D > -75.367339	**Relaxing**
6	IF G > 2.024609 AND E > 3.112653	**Happy**

Phase (3): The last step is to build multi-label classifier based on correlations among labels and rules discovered from applying a rule based algorithm on the transformed dataset. Table10 summarizes some of the multi-label rules discovered from "Emotions" dataset.

TABLE X. MULTI-LABEL RULES DISCOVERED FROM "EMOTIONS" DATASET

Rule #	Multi-Label Rules	Consequence
1	IF AQ > 0.217678 AND B <= 0.090652 AND V > 0.580398 AND AZ > 3.787686 AND AX > 0.060033 AND BD <= 0.173826	**{Sad, Relaxing}**
2	IF AQ <= 0.215792 AND BJ <= 0.108461 AND J <= 1.021892 AND BO <= 0.066288	**{Angry, Relaxing}**
3	IF AS > 0.206592 AND AI > 0.010202 AND D > -76.700621	**{Amazed, Angry}**
4	IF AS > 0.206592 AND AI > 0.010202 AND B <= 0.191563	**{Quite-Still, Sad}**
5	IF AS > 0.208738 AND B <= 0.119991 AND AP > 0.213677 AND BN <= 102 AND D > -75.367339	**{Relaxing}**
6	IF G > 2.024609 AND E > 3.112653	**{Happy, Relaxing}**

To illustrate how this step is performed, let us give a sample rule from the rules set that are obtained after applying PART algorithm on the transformed dataset. The sample rule is:

IF AQ > 0.217678 AND B <= 0.090652 AND V > 0.580398 AND AZ > 3.787686 AND AX > 0.060033 AND BD <= 0.173826 THEN Sad.

Using Association rules among labels that have been discovered earlier, and since there is a rule indicates that (*IF Sad THEN Relaxing*), the rule is rebuilt from the rule based classifier as following:

IF AQ > 0.217678 AND B <= 0.090652 AND V > 0.580398 AND AZ > 3.787686 AND

AX > 0.060033 AND BD <= 0.173826 THEN {Sad, Relaxing}

We repeat the previous process for all rules extracted from the rule based classifier and using the association rules discovered in the first step. The outcome will be the complete set of multi-label rules, which will be used to classify the test instances.

V. Experiments and Results

In this paper, we used two different application domains data sets which they are: Biological, and Musical. For each application domain, one multi-label dataset has been used, as shown in Table11. The datasets are available at (http://mulan.sourceforge.net/datasets.html). The first dataset is called "*Emotions*" and it is concerned about songs according to the emotions they evoke. This data set contains six labels, with label cardinality (LC) equal to 1.869and label density (LD) equal to 0.311. There are 27 distinct label-sets (DLS) in a total number of 593 examples in this dataset. As mentioned earlier, label cardinality (LC) is the average number of labels per example; while label density is the same number (LC) divided by number of labels in the dataset (6 in the emotion dataset as an example).

The second dataset is called "*Yeast*" which is concerned about protein function classification. This dataset contains 2417 examples with 198 distinct label-sets. The Yeast dataset has 14 different labels with cardinality equals to 4.327 and density equals to 0.303.

TABLE XI. Multi-Label Datasets Statistics

Dataset	Domain	# of Instances	# Attributes	# of Labels	DLS	LC	LD
Emotions	Music	593	72	6	27	1.869	0.311
Yeast	Biology	2417	103	14	198	4.327	0.303

Based on the statistics presented in Table14, we are more interested in LC to determine the association's cardinality which is equal to Label Cardinality – 1. Table6 and Table12 summarize the labels that could be found in the datasets which will be used in the evaluation process and the frequency of each label.

TABLE XII. Frequency of "Yeast" Dataset Labels (C1 – C14)

Label	C1	C2	C3	C4	C5	C6	C7
Frequency	762	1038	983	862	722	597	428
Label	**C8**	**C9**	**C10**	**C11**	**C12**	**C13**	**C14**
Frequency	480	178	253	289	289	1799	34

An extensive evaluation process has been made using three evaluation measures, five problem transformation methods, and two algorithm adaptation methods. All multi-label classification methods and all supervised learning algorithms which are used in this paper are implemented using Mulan tool [13] [14] which is a WEKA-based Java package for multi-label classification. All experiments were conducted using the *10-fold* cross validation method. The proposed approach is evaluated using different evaluation measures which are: Accuracy, Hamming Loss, and Harmonic Mean (F1 Measure).

A. Experiments on "Emotions" Dataset

- **Accuracy:** In term of accuracy and as noticed from Fig.3, the proposed approach has the highest accuracy (0.767) among all the multi-label classification methods. The second best accuracy is 0.592 achieved by RAKEL. This indicates that using correlations among labels increase accuracy in a great way.
- **Hamming Loss:** As notices from Fig.4, the proposed approach has the lowest Hamming Loss (0.155) among all the multi-label classification methods. The second best hamming lost is achieved by RAKEL method (0.186), which indicates that the proposed approach decreases both incorrect labels classification and missing labels classification in a good way.
- **The Harmonic Mean (F1 Measure):** As noticed from Fig.5, the proposed approach has the highest Harmonic Mean (0.837) among all multi-label classification methods.

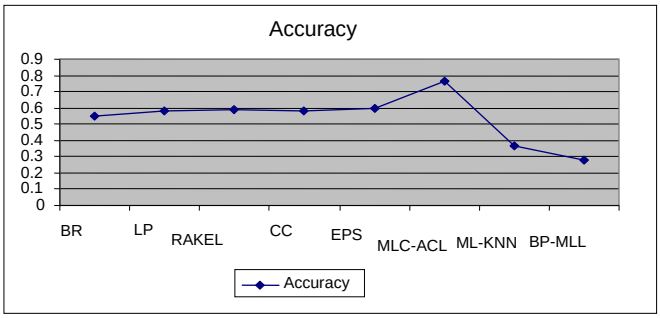


Fig. 3. Difference in accuracy between the proposed approach ((MLC-ACL) and other methods

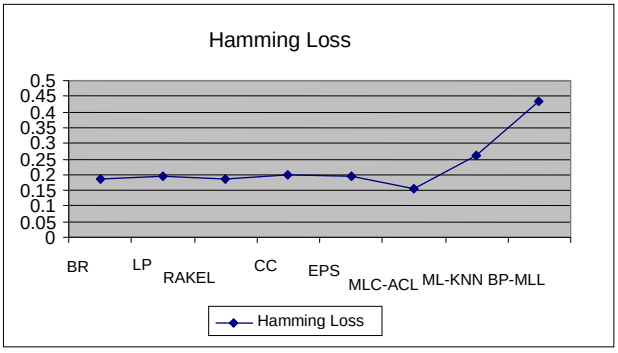


Fig. 4. Difference in Hamming Loss between the proposed approach (MLC-ACL) and other methods

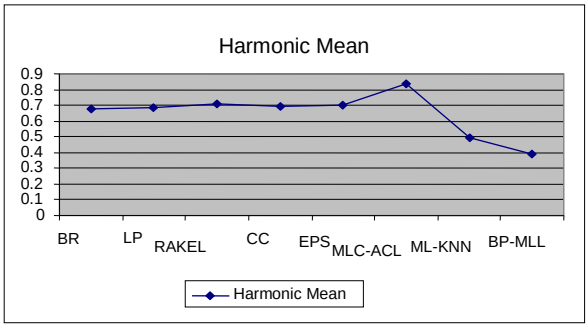


Fig. 5. Difference in Harmonic Mean between the proposed approach (MLC-ACL) and the other methods

B. Experiments on "Yeast" Dataset

Table13 contains the best correlations among labels after applying Predictive Apriori on "Yeast" dataset. Table14 summarizes the results of the evaluation measures on "Yeast" dataset. Table14 shows that the proposed approach has the highest accuracy (0.554), and EPS method has the second highest accuracy (0.537). The proposed approach has the best value for Hamming loss (0.161), while BR and ML-KNN have the second best value (0.193). Finally, the proposed approach has the best value (0.672) of Harmonic mean measure, and ML-KNN has the second best value (0.654) of Harmonic mean.

TABLE XIII. POSITIVE ASSOCIATION RULES USING THE "YEAST" DATASET

Rule #	Rule	Accuracy
1	IF C1 THEN C2	0.49
2	IF C2 THEN C12	0.43
3	IF C3 THEN C12	0.50
4	IF C4 THEN C12	0.51
5	IF C5 THEN C12	0.53
6	IF C6 THEN C12	0.54
7	IF C7 THEN C8	0.63
8	IF C8 THEN C13	0.50
9	IF C9 THEN C8	0.81
10	IF C10 THEN C11	0.82
11	IF C11 THEN C12	0.76
12	IF C12 THEN C12	1.00
13	IF C13 THEN C12	0.80
14	IF C14 THEN C4	0.99

TABLE XIV. EVALUATION RESULTS USING THE "YEAST" DATASET

Method	Accuracy	Hamming Loss	Harmonic Mean
BR	0.522	0.193	0.652
LP	0.530	0.206	0.643
RAKEL	0.493	0.207	0.559
CC	0.521	0.211	0.633
EPS	0.537	0.207	0.654
MLC-ACL	**0.554**	**0.161**	**0.672**
ML-KNN	0.520	0.193	0.654
BP-MLL	0.185	0.322	0.210

VI. CONCLUSIONS AND FUTURE WORK

In this paper, we have investigated the problem of multi-label classification, and the benefits from having the correlations among label in building multi-label rules. The outcome of this research is an algorithm for multi-label classification based on correlations among labels. Unlike previous approaches, this algorithm combines between problem transformation methods with the criteria of selecting least frequent label and unsupervised learning method (Predictive Apriori). The main contributions of this research can be summarized as follows:

- Merging between two different learning tasks, the first task is an unsupervised learning task, which is the task of finding positive association among labels. The second task is a supervised learning task, which is the task of applying any rule-based classifier on the transformed dataset.
- Getting benefits from finding the correlations among labels, in the process of generating multi-label rules. Transforming multi-label dataset into single label dataset causes too loss in information, and by finding correlations among labels, the proposed approach tries to substitute this information loss.
- The proposed approach has much flexibility, since any rule-based classifier could be used in the process of classifying the transformed data set.

As a future work, we suggest Proposing New Problem Transformation Method based on Accuracy of correlations among labels We may adapt the proposed model as following:

- Step1: Discovery of positive correlations among labels
- Step2: Apply problem transformation method based on correlations among labels and using the highest accuracy criteria, which means to select the label that produces the highest accuracy as being antecedent of the association rule.
- Step3: Applying a rule based classifier on the transformed data set and producing the rules set.
- Step4: Generating the multi-label rules set, using the single rules set produced by the classifier in step 3, and the associative rules for each instance that has been discovered in step 1.

Experiment on "Emotions" dataset shows that the adapted model is promising and need to be studied more. When applying the adapted model in "Emotions" dataset, the accuracy was (0.752) which is really close to the accuracy of the proposed model (0.767).

ACKNOWLEDGMENT

Raed Alazaidah thanks all the professors in faculty of information technology in Philadelphia university, especially prof. Said Ghoul. And deep thanks to my best friend Naela Alsalman.

REFERENCES

[1] Thabtah, F., Cowling, P. & Peng, Y, " MMAC: A New Multi-Class, Multi-Label Associative Classification Approach". In Proceedings of the 4th IEEE International Conference on Data Mining (ICDM'04), Brighton, UK, 2004, pp. 217–224.

[2] Neda Abdelhamid, Aladdin Ayesh, Fadi Thabtah, Samad Ahmadi, Wael Hadi," MAC: A multiclass associative classification algorithm", Journal of Information & Knowledge Management , volume 11 , ssue 2, June 2012, pp. 1250011-1 - 1250011-10.

[3] Boutell, M., Luo, J., Shen, X., Brown, C, "Learning Multi-label Scene Classification". PatternRecognition, 2004, 37:1757–1771.

[4] Tsoumakas, G., Vlahavas, I, "Random k-labelsets: An ensemble method for multilabel classification". In: Proceedings of the 18th European Conference on Machine Learning (ECML 2007), Warsaw, Poland, 2007,pp. 406–417.

[5] Read, J., Pfahringer B., Holmes G, "Multi-Label Classification using Ensembles of Pruned Sets", 8th IEEE International Conference on Data Mining, 2008.

[6] F¨urnkranz J. and H¨ullermeier E, "Pairwise Preference Learning and Ranking", In N. Lavra˘c,D. Gamberger, H. Blockeel, and L. Todorovski, (eds), the 14th EuropeanConference on 11Machine Learning (ECML-03,) Cavtat, Croatia, Lecture Notes in Artificial Intelligence, Springer-Verlag, 2837, 2003, pp. 145–156.

[7] Jabez Ch, "A Statistical Approach for Associative Classification", European Journal of Scientific Research, 58(2), 2011, pp.140-147.

[8] Read, J., Pfahringer, Bernhard, Holmes, Geo_rey,and Frank, Eibe." Classifier chains for multi-label Classification". In Proceedings of the European Con-ference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases, 2009, pp 254-269.

[9] Clare, A., King, R ," Knowledge discovery in multi-label phenotype data". In Proceedings of the 5th European Conference on Principles of Data Mining and Knowledge Discovery (PKDD 2001), Freiburg, Germany pp. 42–53.

[10] Zhang, M. L., Zhou, Z. H , "ML-kNN: A Lazy Learning Approach to Multi-Label Learning", Pattern Recognition, 2007, 40, pp. 2038–2048.

[11] Scheffer, T., "Finding Association Rules that Trade Support Optimally Against Confidence", In Proceedings of the 5th European Conference on Principles and Practice of Knowledge Discovery in Databases (PKDD'01), Freiburg, Germany: Springer-Verlag, 2001, pp. 424-435.

[12] Neda Abdelhamid, Fadi Thabtah, " Associative Classification Approaches: Review and Comparison", Journal of Information & Knowledge Management, 2014 .

[13] Tsoumakas G. Spyromitros-Xioufis E. Vilcek J. Vlahavas I., "MULAN: A Java Library for Multi-Label Learning", Journal of Machine Learning Research 12, 2011, pp. 2411-2414.

[14] Neda Abdelhamid, Aladdin Ayesh, Fadi Thabtah, " Phishing detection based Associative Classification data mining", Expert Systems with Applications, Volume 41, Issue 13 , 2014,pp. 5948-5959

Developing Software Bug Prediction Models Using Various Software Metrics as the Bug Indicators

Varuna Gupta
Research Scholar, Christ University, Bangalore

Dr. N. Ganeshan
Director, RICM, Bangalore

Dr. Tarun K. Singhal
Dean-Academic, INMANTEC, Gzb

***Abstract*—The bug prediction effectiveness reasonably contributes towards enhancing quality of software. Bug indicators contribute significantly in determining the bug prediction approaches and help in achieving software reliability. Various comparative research studies have indicated that Depth of Inheritance (DIT), Weighted Method per Class (WMC), Coupling between Objects (CBO) and Lines of Code (LoC) have significantly established themselves as reliable bug indicators for comprehensive bug predictions.**

The researchers have carried out a quantitative research and have developed prediction models using above bug indicators as models input and have applied these models on open source projects (Camel and Ant). During this research, the results demonstrates that there is significant correlation between size oriented metrics (bug indicators) such as DIT, WMC, CBO, LoC and bugs. Overall, DIT takes dominance in achieving better impact on predicting bugs than WMC, CBO and LoC.

The outcomes of the present research study would be of significance to software quality practitioners worldwide and would help them in prioritizing the efforts involved in bug prediction.

Keywords—Bug Prediction; DIT; WMC; CBO; LoC; SRGM

I. INTRODUCTION

Software reliability is considered critical and important aspect of software quality. Organizations pay due emphasis in detecting the quality of software product at an early stage to avoid late embarrassments arising due to late detection culminating in poor quality product ultimately. This approach ensures that organizations are able to redesign wherever possible and ensure consistent quality throughout. Organizations aim to ensure savings towards costs of development, reduction in time to develop and high reliability of software products.

Various attributes such as proneness to faults, testing efforts, maintenance efforts etc govern the quality of software products. Through this research, we have considered proneness to bugs as bug predictor utilizing DIT, WMC, CBO and LoC indicators within the realm of this research.

Various bug indicators proposed during last few decades have made the selection of right bug indicator a demanding task considering the complexity and nature of varying software development processes. In the wake, a number of researchers have predominantly proposed product oriented bug indicators. The testers across many organizations dedicate time and resources by allocating same priorities across all components of a project, which is not considered as an optimal approach.

Parts of the software systems don't have uniformity in bug distribution. This calls for comprehensive identification of files containing bugs throughout the project. The testers with such knowledge would be able to identify and prioritize the appropriate tests while achieving efficiency in testing process. In order to achieve the said, it is essential to ensure availability of appropriate software bug prediction models. The main objective of this research is to construct software bug prediction models using four bug indicators as the model input. The metrics collected by promise repository are used as the model input. Therefore, the model construction process allows assessment of appropriateness of the collected metrics as usable bug predictors. The predicted number of bugs for the files is the model output.

The present research has been organized into six sections. Section I introduces the concepts and practices being adopted in software bug prediction. Section II contains detailed review of literature. Section III demonstrates the process map adopted by the researchers. Section IV proposes modeling framework. Section V & VI contain analysis, conclusion and future research work.

Need of the Study

The generic realization is that software practitioners need to focus early on bug prediction approaches to ensure reasonable quality in software products. Therefore, a comprehensive research was needed to widen the scope of bug prediction approaches and identify bug indicators causing significant impact on software quality.

Objectives

1) To assess the correlation of bug indicators (DIT, WMC, CBO, LoC) with software bugs.

2) To develop software bug prediction models using bug indicators (DIT, WMC, CBO, LoC) as model inputs.

3) To compare the relative effectiveness of DIT, WMC, CBO and LoC towards prediction of bugs in Camel and Ant projects.

II. RELATED WORK

A significant amount of work has been cited using product metrics to predict bug prone files. Though major work has utilized Chidamber and Kemerer (CK) metrics suite [18] to

predict accurately pre and post release bugs in commercial and open source systems [23, 10, 8, 20, 13, 22]. Further, though CK metrics suite, empirical justification has also been made regarding usefulness in bug prediction [3, 6, 14].

Pareto analysis has also been used for evaluating the ability of models for identification of fault-prone classes, modules and files. As substantiated with presence of 80% of bugs in 20% of files [15, 26, 24, 25].

Linear regression has been widely considered as a common technique for bug prediction. Also DIT has been demonstrated to carry a linear relationship with bugs [16]. Further, our data was linear in nature advocating application of linear regression. Still, keeping with [1], which suggested application of nonlinear regression as better indicator for this type of data, so decided to go ahead with non linear regression.

Logistic regression models have also been used to identify fault-prone modules [4]. CK metrics suite was also used to find fault-prone classes [19]. This work involved investigation of two C++ written projects and followed with outcome involving analysis of 43-48% of classes to cover for 80% of the bugs

Bug prediction models were created based on the module size representing Line of Code (LoC). The models produced outputs in strong correlation with actual data [12]. These models suggested considering LoC in the bug prediction models.

A majority of CK metrics were found to be effective predictors for fault-proneness of class. In addition, DIT and Response for a Class (RFC) were found to be carrying more influence on the dependent variable [2].

A study on data from an industrial system comprising of more than 200 C++ subsystems added different metrics than CK metrics and applied logistic regression to evaluate those metrics. The outcomes suggested WMC and DIT as significant indicators for finding fault-prone classes [21].

Another research applying logistic regression on data from a telecommunication system having 174 C++ classes demonstrated close association of WMC, RFC and Coupling between Objects (CBO) with software bugs [5]. Another research using univariate logistic regression also identified WMC and SLOC as significant predictors [11].

Another research using data from two commercial applications, one having 150 classes and 23 KSLOC while other having 144 classes and 25 KSLOC evaluated the influence of six CK metrics on the number of bugs and identified RFC and DIT as most significant variables [19].

As per recent citations of the research works carried out, no significant amount of work has been done on the use of logistic reliability growth model for bug prediction.

Proneness to Bugs

Software failing to fulfill the specified requirement needs to be fixed. Signifying that the mistake has been committed between the initial requirement and the final operation of the software system. Since source code matters the most corresponding to the realization of the software system, the errors in source code are called bugs. There are changes that error may not become a bug. However, we need to fix it if it ultimately becomes a bug causing a failure. The proneness of bugs depends on reasons like DIT, WMC, CBO , LoC.,

DIT (Depth of Inheritance Tree): The maximum length from the root to a given class in the inheritance hierarchy. DIT is defined as the maximum length inheritance path from the class to the root class [19].

WMC (Weighted Methods per Class): WMC is defined as the sum of the complexity of the methods of the class. It is equal to the number of methods when all methods are of the complexity equal to UNITY. The sum of normalized complexity of every method in a given class.

CBO (Coupling Between Objects): The CBO metric represents the number of classes coupled to a given class. These couplings can occur through method calls, field accesses, inheritance, method arguments, return types and exceptions [18].

LOC (Line of Code): the LOC metric based on Java binary code represents sum of number of fields, number of methods and number of instructions in every method of the investigated class.

III. PROCESS MAP

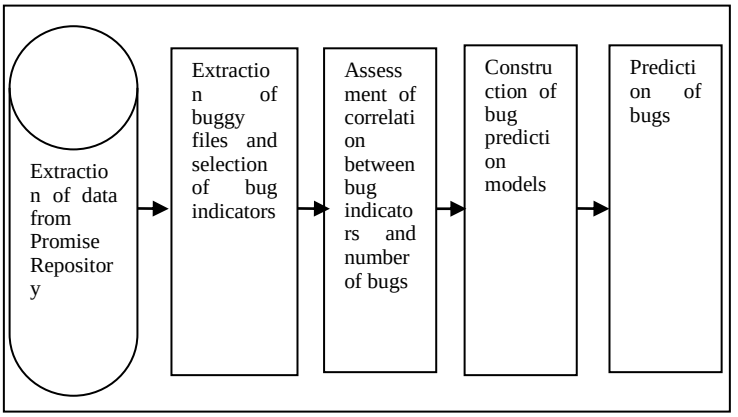


Fig. 1. Process Map

In this paper, the proposed process map is using the mixed method combining qualitative and quantitative research methods. The research work is detailed in five phases as shown in Figure 1.

A. Extraction of Data: Researchers have used PROMISE repository to extract the bug indicators (DIT, WMC, CBO and LoC) and bug data. The reason for selecting the open source projects from PROMISE repository was that it is a trustworthy software foundation having positive feedback from software users. It is also well-recognized in the software community.

B. *Extraction of buggy files and selection of bug indicators:* Two open source projects (Camel and Ant) were preferred to extract bug data from and selection of bug indicators for the analysis. Proper literature review was performed to select suitable bug indicators (DIT, WMC, CBO and LoC) for this research.

C. *Assessment of correlation between bug indicators and bugs:* Pearson correlation analysis was performed to assess the correlation between the various bug indicators (DIT, WMC, CBO, LoC) and number of bugs.

D. *Construction of prediction models:* After significant correlation between bug indicators and bugs, researchers have constructed prediction models using logistic software reliability growth model on extracted data from PROMISE bug database.

E. *Prediction:* After successful conclusion of the above four sub processes, finally predicted bugs was given as the model output.

IV. MODELLING FRAMWORK

A. Software Reliability Growth Models (SRGM)

Software reliability growth models are a statistical exclamation of detected bug's data using various mathematical functions. To predict the number of bugs in the code these mathematical functions are used. There are many types of software reliability growth models as to predict future bugs or failure rates.

B. Models Assumptions

Some of the general assumptions (apart from some special ones for specific models discussed) for the above model are as follows:

a) Software system is subject to failure during execution caused by bugs remaining in the system.

b) Failure rate of the software is equally affected by bugs remaining in the software.

c) The number of bugs predicted at any time instant is proportional to the actual number of bugs in the software.

d) Bug indicators referring the software size and its proportional impact on bugs have the capabilities of certain prediction.

e) All bugs are mutually independent from bug prediction point of view.

f)Bug prediction rate/bug detection rate is a logistic learning function as it is expected the learning process will grow with time.

g) The bug prediction phenomenon is modeled by Non Homogeneous Poisson Process (NHPP).

C. Models Notations

a- initial fault-content of the software.

k- A constant parameter in the logistic learning function

b_1- bug prediction rate/detection rate per unit time.

M (t) - expected number of bugs predicted.

Bug prediction models using SRGM are given by:

$$m(t) = a / \left(1 + k.e^{-b_1.t}\right) \qquad (4.1)$$

Prediction model-1

DIT is considered as a first model input referring to the below mentioned proposed model:

$$m(t) = a / \left(1 + k.e^{-b_1.dit}\right) \qquad (4.2)$$

Prediction model-2

WMC is defined as a second model input referring to the below mentioned proposed model:

$$m(t) = a / \left(1 + k.e^{-b_1.wmc}\right) \qquad (4.3)$$

Prediction model -3

CBO is defined as a third model input referring to the below mentioned proposed model:

$$m(t) = a / \left(1 + k.e^{-b_1.cbo}\right) \qquad (4.4)$$

Prediction model -4

LoC is defined as a fourth model input referring to the below mentioned proposed model:

$$m(t) = a / \left(1 + k.e^{-b_1.loc}\right) \qquad (4.5)$$

D. Goodness of Fit Criteria

The performance of a bug prediction model is judged by its ability to fit the past software reliability data and to predict satisfactorily the future behavior from present and past data behavior. The following criteria defined as:

1) *Coefficient of Multiple Determinations (R^2)*
2) *Bias*
3) *Variation*
4) *The Root Mean Square Prediction Error (RMSPE)*
5) *Mean Square Error (MSE)*

Bug Prediction Parameter Estimation

To examine the effectiveness of software bug prediction models using four indicators as model input, a set of comparison criteria is used to compare models quantitatively. The different comparison criterions used in our paper are as follows:

1) *Coefficient of Multiple Determination (R^2):*

This Goodness-of-fit measure has been used to investigate significance in trend existing in prediction of bugs. This coefficient was used as the ratio of the Sum of Squares (SS) derived from the trend model to that from a constant model subtracted from 1, that is

$$R^2 = 1 - \frac{residual \quad SS}{corrected \quad SS}$$

R^2 measures the percentage of the total variation about the mean accounted for by the fitted curve. It ranges in value from 0 to 1. Small values indicate that the model does not fit the data well. With movement of value towards 1, the model significantly explains the variation in the data [7].

2) Bias: The difference between the actual and predicted number of bugs at any instant of time i is known as Prediction Error (PE_i). The average of PEs is known as bias. With movement of value towards 0, the model significantly explains low presence of prediction error. The bias is defined mathematically as [9]:

$$Bias = \frac{\sum_{i=1}^{k}\left(\hat{m}(t_i) - m_i \right)}{k}$$

Where m_i indicates actual bugs, m(t) indicates predicted bugs and k is the number of observations in the data set.

3) Variance: The variance is defined as [9].

$$Variance = \sqrt{\frac{1}{k-1}\sum_{i=1}^{k}\left(m_i - \hat{m}(t_i) - Bias \right)^2}$$

4) Root Mean Square Prediction Error (RMSPE): It measures the closeness with which the model predicts the bugs and mathematical representation of this characteristic is given as [9].

$$RMSPE = \sqrt{Variance^2 + Bias^2}$$

5) Mean Square Error (MSE): MSE measures the difference between the predicted and actual values of bugs, and is given mathematically as **[**17**].**

$$MSE = \frac{\sum_{i=1}^{k}\left(m_i - \hat{m}(t_i) \right)^2}{k - p}$$

Where k is the number of observations in the data set and p is the number of parameters.

E. Data Sets

The data about bug indicators and bugs has been collected from PROMISE repositories. The following data sets have been used with explanations marked in:

Data Set 1(Camel) Apache Camel is a powerful open source integration framework based on known Enterprise Integration Patterns with powerful Bean Integration.

Data Set 2 (Ant) Ant is a well known Java-based, shell independent build tool.

V. Analysis and Conclusion

While checking the accuracy of different proposed models of bug prediction using different bug indicators, researchers have first estimated the unknown parameters of bug data for final software product on bug cumulative consumption data. Then, to judge the fitting of various proposed models of prediction given by equations (4.2), (4.3) (4.4) and (4.5) R^2, bias, variation, RMSPE and MSE have been calculated as the performance measures. Table I and Table II depict the estimated values for the parameters while Table III provides the correlation criteria and finally Table IV and Table V summarizes the estimated and optimized values of attributes of proposed models.

TABLE I. Estimated Parameters of Proposed Models Using Ds-1

S. No.	Parameters	Estimated parameters values			
		DIT	***WMC***	***LOC***	***CBO***
1	a	136.41	139.99	135.89	161.86
2	K	24.48	10.16	12.57	11.86
3	b_1	.071	.008	.001	.006

TABLE II. Estimated Parameters of Proposed Models Using Ds-2

S. No.	Parameters	Estimated parameters values			
		DIT	WMC	LOC	CBO
1	a	51.09	46.32	48.19	46.59
2	K	11.54	12.12	18.07	14.11
3	b_1	.046	.013	.001	.015

In our research, researchers observed significant correlations of WMC, DIT, CBO and LOC with bugs. In this research only highly correlated four metrics have shown from each data set that are listed in Table III. The interesting part of this result is that all four indicators are correlated significantly with software bugs.

TABLE III. Correlation Table

Project	Metrics	Correlation with Bugs
Camel	DIT	.976
	WMC	.987
	LOC	.984
	CBO	.992
Ant	DIT	.997
	WMC	.989
	LOC	.992
	CBO	.991

TABLE IV. Estimated and Optimal Values of Attributes for Four Prediction Models For Ds-1

Project	Metrics	R2	Bias	Variance	RMSE	MSE
Camel	DIT	99.5	-0.271	3.318	3.329	11.253
	WMC	98.9	0.183	4.712	4.716	23.048
	LOC	98.9	0.122	5.518	5.519	22.687
	CBO	98.6	0.141	5.271	5.273	28.88

TABLE V. Estimated and Optimal Values of Attributes for Four Prediction Models For Ds-2

Project	Metrics	R2	Bias	Variance	RMSE	MSE
Ant	DIT	99.1	0.089	1.349	1.352	1.893
	WMC	98.3	0.156	1.891	1.898	3.693
	LOC	98.9	0.132	1.505	1.511	2.333
	CBO	98.9	0.147	1.507	1.514	2.331

In table III researchers observed significant correlations of WMC, DIT, CBO and LOC with bugs. Table IV depicted that in case of DS-1 using prediction model 4.2 the predictive model coefficient of determination is 0.995 it means 99.5% of the variation in bugs is associated with number of predictor. Whereas using model 4.3, model 4.4 and model 4.5 the variation in bugs is 98.9%, 98.6% and 98.9% respectively.

Table V depicted that in case of DS-2 using prediction model 4.2 the predictive model coefficient of determination is 0.991 it means 99.1% of the variation in bugs is associated with number of predictor. Whereas using model 4.3 model 4.4 and model 4.5 the variation in bugs is 98.3%, 98.3% and 98.9% respectively.

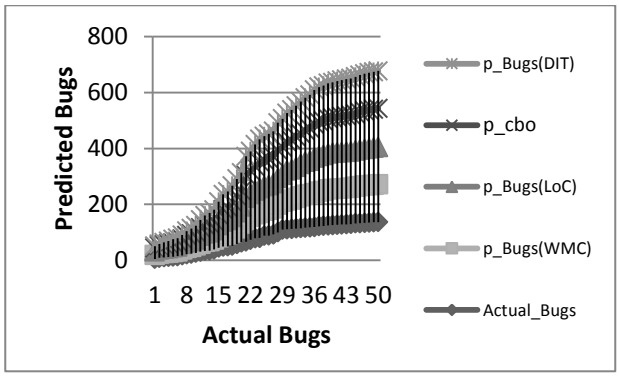


Fig. 2. Graph for Pattern of Actual and Predicted Software Bugs of DS-1

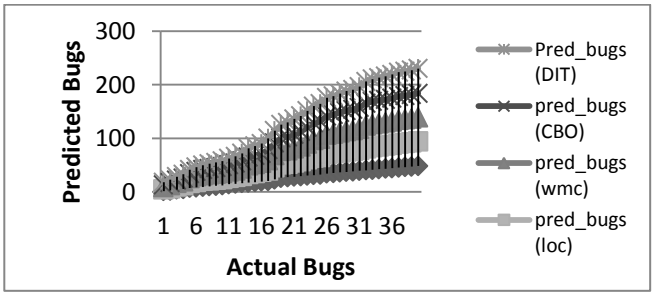


Fig. 3. Graph for Pattern of Actual and Predicted Software Bugs of DS-2

As shown above graphs in Figure – 2 and Figure – 3, the predicted number of bugs is significantly higher than actual number of bugs.

The research has comprehensively designed and tested four models using DIT, WMC, CBO and LoC as model inputs. These models produced significant results on all four model inputs. However, model using DIT as input was shown to be better performing than the other three models. This conclusion can serve as strong motivation for software practitioners to prioritize and allocate sufficient resources towards DIT because of its better performance in comparison to WMC, CBO and LoC.

VI. Future Work

More product metrics as bug indicators can be included in future research work. More open source data sets can also be included to bring higher reliability in bug prediction. An effort can be made of applying different non linear regression models on same two data sets already considered in present research work.

References

[1] A. Bernstein, J. Ekanayake, and M. Pinzger. Improving defect prediction using temporal features and non linear models. In Proc. Int'l Workshop on Principles of Softw. Evolution, pages 11–18, 2007.

[2] B asili , V.R., L.C. Briand and W.L. Melo, 1996. A validation of object-oriented design metrics as quality indicators. IEEE Trans. Software Eng., 22: 751-761. DOI10.1109/32.544352.

[3] CATAL C., DIRI B. and OZUMUT B., An Artificial Immune System Approach for Fault Prediction in Object-Oriented Software. Proc. of Dependability of Computer Systems, 2007, 238-245.

[4] DENARO G. and PEZZE M., An Empirical Evaluation of Fault-Proneness Models. Proc. of International Conference on Software Engineering (ICSE), 2002.

[5] El Emam, K., S. Benlarbi, N. Goel and S.N. Rai, 2001. The confounding effect of class size on the validity of object-oriented metrics. IEEE Trans. Software Eng., 27: 630-650. DOI: 10.1109/32.935855.

[6] JURECZKO M., Use of software metrics for finding weak points of object oriented projects. Proc. Of Metody i narzędzia wytwarzania oprogramowania 133-144, 2007 (in Polish).

[7] K. C. Chiu, Y. S. Huang, and T. Z. Lee, "A study of software reliability growth from the perspective of learning effects," Reliability Engineering and System Safety, pp. 1410–1421, 2008.

[8] K. E. Emam, W. Melo, and J. C. Machado. The prediction of faulty classes using object-oriented design metrics. J. Syst. Softw., 56(1):63–75, 2001.

[9] K. Pillai and V. S. S. Nair, "A model for software development effort and cost estimation," IEEE Trans. Softw. Engineering, vol. 23, no. 8, pp. 485–497, 1997

[10] L. C. Briand, J. W. Daly, and J. K. W¨ust. A unified framework for coupling measurement in object-oriented systems. IEEE Trans. Softw. Eng., 25(1):91– 121, 1999.

[11] Malhotra R., and Jain A., "Fault prediction using statistical and machine learning methods for improving software quality", Journal of Information Processing Systems, Vol.8, No.2, June 2012

[12] MENDE T., KOSCHKE R., Revisiting the Evaluation of Defect Prediction Models. Proc. of PROMISE, 2009.

[13] N. Nagappan, T. Ball, and A. Zeller. Mining metrics to predict component failures. In Proc. Int'l Conf. on Softw. Eng. (ICSE'06), pages 452–461, 2006.

[14] OLAGUE H. M., ETZKORN L. H., GHOLSTON S. and QUATTLEBAUM S., Empirical Validation of Three Software Metrics Suites to Predict Fault-Proneness of Object-Oriented Class Developed Using Highly Iterative or Agile Software Development Processes. IEEE Trans. on Software Engineering, 33(6), 2007, 402-419.

[15] OSTRAND T. J., WEYUKER E. J. and BELL R. M., Predicting the Location and Number of Faults in Large Software Systems. IEEE Trans. on Software Engineering, 31(4), 2005, 340-356.[8]

[16] S. Dowdy, S. Weardon, and D. Chilko. Statistics for Research. Probability and Statistics. JohnWiley and Sons, Hoboken, New Jersey, third edition, 2004.

[17] S. Hwang and H. Pham, "Quasi-renewal time-delay fault-removal consideration in software reliability modelling," IEEE Trans. Systems, Man and Cybernetics-Part A: Systems and Humans, vol. 39, no. 1, January 2009.

[18] S. R. Chidamber and C. F. Kemerer. A metrics suite for object oriented design. IEEE Trans. Softw. Eng., 20(6):476–493, 1994.[1]

[19] SUCCI G., PEDRYCZ W., STEFANOVIC M. and MILLER J., Practical assessment of the models for identification of defect-prone classes in object-oriented commercial systems using design metrics. Journal of Systems and Software 65(1), 2003, 1-12.

[20] T. Gyimothy, R. Ferenc, and I. Siket. Empirical validation of object-oriented metrics on open source software for fault prediction. IEEE Trans. Softw. Eng., 31(10):897–910, 2005.

[21] Tang, M.H., M.H. Kao and M.H. Chen, 1999. An empirical study on object-oriented metrics. Proceedings of the 6th International Symposium

on Software Metrics, Oct. 04-06, IEEE Computer Society, Boca Raton, FL., USA., pp: 242-249. DOI: 10.1109/METRIC.1999.809745.

[22] T. Zimmermann, R. Premraj, and A. Zeller. Predicting defects for Eclipse. In Proc. Int'l Workshop on Predictor Models in Software Engineering (PROMISE'07), pages 1–7, 2007.

[23] V. R. Basili, L. C. Briand, and W. L. Melo. A validation of object-oriented design metrics as quality indicators. IEEE Trans. Softw. Eng., 22(10):751–761, 1996.

[24] WEYUKER E. J., OSTRAND T. J. and BELL R. M., Adapting a Fault Prediction Model to Allow Widespread Usage. Proc. of PROMISE, 2006.

[25] WEYUKER E. J., OSTRAND T. J. and BELL R. M., Do too many cooks spoil the broth? Using the number of developers to enhance defect prediction models. Empirical Software Engineering, 13(5), 2008, 539-559.

[26] WEYUKER E. J., OSTRAND T. J. and BELL R. M., Comparing Negative Binomial and Recursive Partitioning Models for Fault Prediction. Proc. of PROMISE, 2008.

The Effects of Different Congestion Management Algorithms over Voip Performance

Szabolcs Szilágyi
Faculty of Informatics
University of Debrecen
Debrecen, Hungary

***Abstract*—This paper presents one of the features of DS (Differentiated Services) architecture, namely the queuing or congestion management. Packets can be placed into separate buffer queues, on the basis of the DS value. Several forwarding policies can be used to favor high priority packets in different ways. The major reason for queuing is that the router must hold the packet in its memory while the outgoing interface is busy with sending another packet. The main goal is to compare the performance of the following queuing mechanisms using a laboratory environment: FIFO (First-In First-Out), CQ (Custom Queuing), PQ (Priority Queuing), WFQ (Weighted Fair Queuing), CBWFQ (Class Based Weighted Fair Queuing) and LLQ (Low Latency Queuing). The research is empirical and qualitative, the results are useful both in infocommunication and in education.**

***Keywords*—CBWFQ; congestion; CQ; FIFO; LLQ; Pagent; PQ; queuing; WFQ**

I. INTRODUCTION

At the beginning computer networks were designed mainly for data transfer such as FTP and email, where delay was considered to be unimportant. In most cases the delivery service was effective, and the TCP protocol dealt with data losses. As the multimedia applications became popular (voice transfer, video conferences), separate telephone and video communication networks were set up (see Fig. 1). Nowadays, office and company networks are transformed into one converged network (see Fig. 2), in which the same network infrastructure is used to ensure all the requested services [1].

Although converged networks have many advantages, there are some disadvantages too, namely the competition for network resources (buffers of routers), which leads to congestion. Delay in delivering the packets, jitter, loss of packets are consequences of congestion. Different applications show different sensitivity to these issues. For example, FTP is not impacted by delay and jitter, whereas the multimedia applications (e.g. interactive voice) are very sensitive to them and the loss of packets too [2]. Quality of Service (QoS) was introduced to handle this problem, and it is able to provide different priority to different applications, users, or data flows, or to guarantee a certain level of performance to a data flow [3], [4].

This research was realized in the frames of TÁMOP 4.2.4. A/2-11-1-2012-0001 „National Excellence Program – Elaborating and operating an inland student and researcher personal support system”. The project was subsidized by the European Union and co-financed by the European Social Fund.

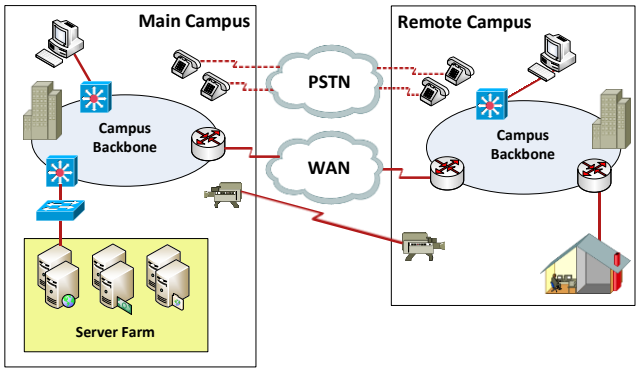


Fig. 1. A classical non-converged network

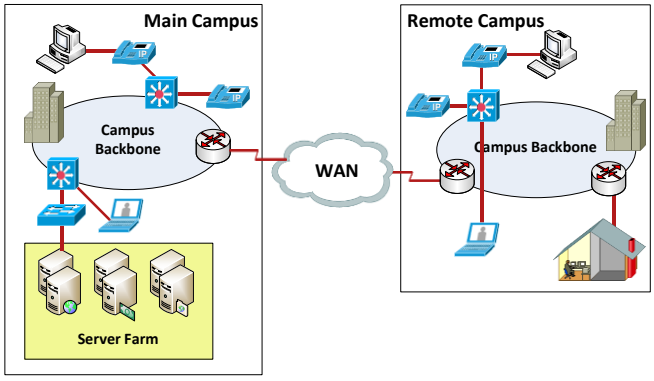


Fig. 2. A converged infocommunication network

In accordance with the QoS requirements and recommendations for the interactive voice traffic packet loss should be no more than 1%, one-way latency should be not exceed 150ms and the average one-way jitter should be targeted at less than 30ms [5].

In the IP header there are some fields which can be used to make distinction between the packets of different applications, for example the Type of Service (ToS) field [6]. Different technics are used for congestion management (PQ, CQ, WFQ, CBWFQ and LLQ). Congestion avoidance (WRED), traffic shaping and traffic policing are also used by the QoS technology in order to control data traffic [7]. This article focuses on the most important component, namely the congestion management.

Our purpose is to analyze and evaluate the efficiency of these congestion management algorithms using a laboratory

environment. This paper examines the following methods: FIFO, PQ, CQ, CBWFQ, WFQ and LLQ. It is important to note that these algorithms have real effect only in the case of congestion.

The network topology for the performance evaluation is identical to the one used in former articles (see e.g. [8]-[10]). This paper continues to study the queuing technologies for congestion management. In [8] and [11] the authors considered three algorithms: FIFO, PQ and WFQ. The conclusion was that WFQ is the most efficient for multimedia applications. In addition to these three new algorithms were investigated: CQ, CBWFQ and LLQ. The main result of this paper is that for multimedia applications (mainly for voice transfer) LLQ is more efficient then WFQ.

The detailed description of the algorithms has been discussed in several papers already (see e.g. [12]-[14]).

II. THE MEASUREMENT ENVIRONMENT

The measurement environment network topology is shown on Fig. 3, which was built at the network laboratory of the Faculty of Informatics, University of Debrecen.

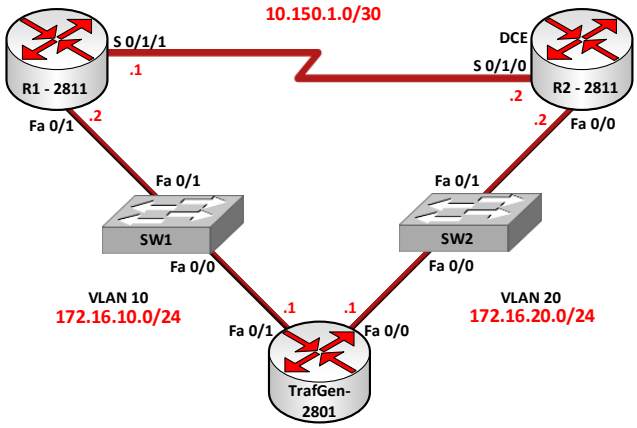


Fig. 3. The measurement environment

The measurement environment consists of three routers (two Cisco 2811 type routers and one Cisco 2801 router) and two switches. The routers are connected with a point-to-point link, having the speed of 128000 cycles per second. The rest of the network devices are connected with 10 Mbps Ethernet links. The part between the two routers is actually a narrow cross-section where congestion can happen. For this reason the congestion management algorithms are activated in this area (see [15]-[16]).

The *Cisco IOS 12.4* operating system was running on the *R1* and *R2* routers, represented on Fig. 3. The *TrafGen* router was responsible for the functioning of the communication endpoints. This was used to generate the traffic, and the operating system run on *TrafGen* was *c2801-tpgen+ipbase-mz.PAGENT.4.3.0* [17], which enabled the traffic generation, attached timestamps to the outgoing packages, and performed the statistical analysis based on the rate of incoming packets.

In order to distinguish between the generated and incoming traffic, two Cisco 2960 switches were used (*SW1*, *SW2*). These created two Virtual LANs, namely *VLAN 10* and *VLAN 20* [18]. A serial connection was created between *R1* and *R2*, which simulated a slow WAN. Three types of traffic were generated, similarly to the previous papers: an FTP, Video and VoIP traffics. In the next section the traffic generation code is shown used by *TrafGen* router.

A. The traffic generation

The following code was used for traffic generation [19]:

```
wait-after-stop 1 ! Waiting time (sec)
!FTP traffic generation
fastethernet0/1 ! The TrafGen router output interface name
add tcp ! Adding a traffic stream (TCP)
datalink ios-dependent fastethernet0/1 ! The output interface name
l2-arp-for 172.16.10.2 ! Layer 2 ARP message to default gateway
l3-src 172.16.10.1 ! Layer 3 source IP address
l3-dest 172.16.20.1 ! Layer 3 destination IP address
l3-tos 0x00 ! Layer 3 packet header ToS byte value
l4-src 21 ! Transport layer source port number
l4-dest 21 ! Transport layer destination port number
name FTP ! Name of the generated traffic
rate 20 ! Setting the packet send rate
length 1434 ! Setting packet length (Bytes)
delayed-start 0 ! Delaying start of packet generation (sec)
send 206 ! Sending packets
fastethernet0/0 ios-dependent capture
! The TrafGen router input interface name
!VIDEO traffic generation
fastethernet0/1
add tcp
datalink ios-dependent fastethernet0/1
l2-arp-for 172.16.10.2
l3-src 172.16.10.1
l3-dest 172.16.20.1
l3-tos 0x22
l4-src 4249
l4-dest 1720
name VIDEO
rate 50
length 1500
burst on ! Sending traffic stream in bursts
burst duration off 1000 to 2000
burst duration on 1000 to 3000
delayed-start 0
send 333
fastethernet0/0 ios-dependent capture
!VOICE traffic generation
fastethernet0/1
add udp
datalink ios-dependent fastethernet0/1
l2-arp-for 172.16.10.2
l3-src 172.16.10.1
l3-dest 172.16.20.1
```

```
l3-tos 0x2E
l4-src 44899
l4-dest 5060
name VOICE
rate 50
length 150
delayed-start 0
send 513
fastethernet0/0 ios-dependent capture
```

B. *The implementation of congestion management algorithms*

The part between the R1 and R2 routers is actually a narrow cross-section where congestion can happen. For this reason the congestion management algorithms were activated in this area (between the *R1*' S 0/1/1 and *R2*' S 0/1/0 interfaces). These codes (for FIFO, PQ, CQ, WFQ, CBWFQ and LLQ) can be found in *APPENDIX*.

III. MEASUREMENT RESULTS

As in the previous works [9]-[10], the length of the measurement time was 5 minutes in each case. The measurements were recorded in every second. Easy to observe in the traffic generation code, that the generated voice traffic average was 513 packets per second. As in previous articles (see e.g. [8]-[11]) the following areas were examined: packet loss, end-to-end delay and jitter (delay variation).

Concerning the packet loss of voice packets (see Fig. 4) LLQ and PQ algorithms have proven to be most effective, followed by the CBWFQ. It can be observed that while the previous works based on simulation results concluded that WFQ was the best congestion management algorithm, our measurement results showed, that the WFQ performance was behind the performance of LLQ, PQ and CBWFQ. The CQ has the next poor algorithm performance, while the least efficient queuing scheduler was the FIFO.

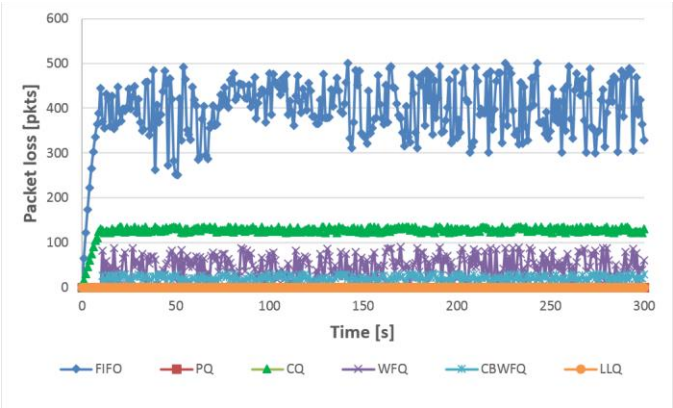


Fig. 4. VoIP packet loss

Fig. 5 shows the same content as Fig. 4, except that the former does not include the efficiency representation of the two least efficient congestion management algorithm. Thus it is prominently observable the difference of performance of PQ, WFQ, CBWFQ and LLQ in packet loss. Easy to see, that in the case of LLQ and PQ there was no packet loss due to the absolute priority queue, in which the real-time voice was classified. Subsequently, CBWFQ performance was the most effective, and finally the WFQ's.

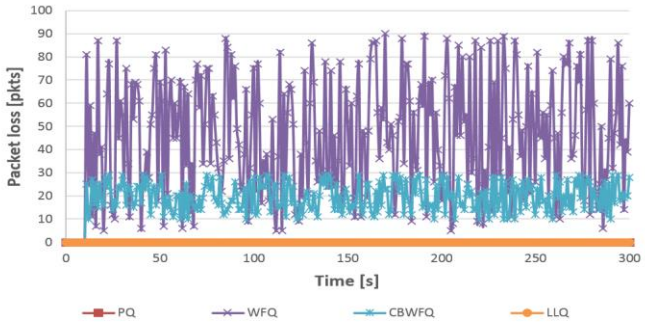


Fig. 5. VoIP packet loss for PQ, WFQ CBWFQ and LLQ

With respect of voice packet delay (see Fig. 6 and Fig. 7), LLQ and PQ algorithms managed to squeeze those values below 100 ms, while the CQ has under 255ms, which already exceeds the threshold set by the QoS requirement. It is clear that in the case of WFQ and CBWFQ the delay is a little more than 1 second, while in the case of FIFO than can reach up to 8.5 seconds, which are unacceptable values provided by the QoS requirements.

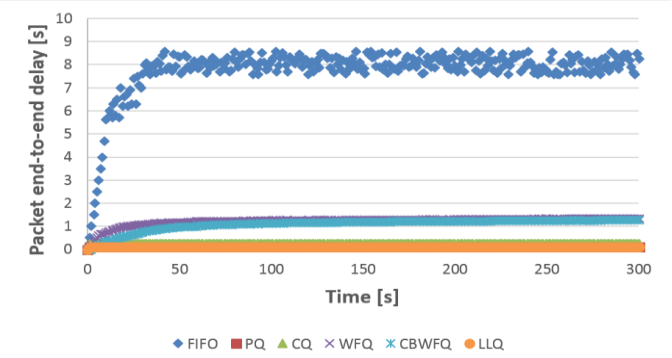


Fig. 6. VoIP traffic delay

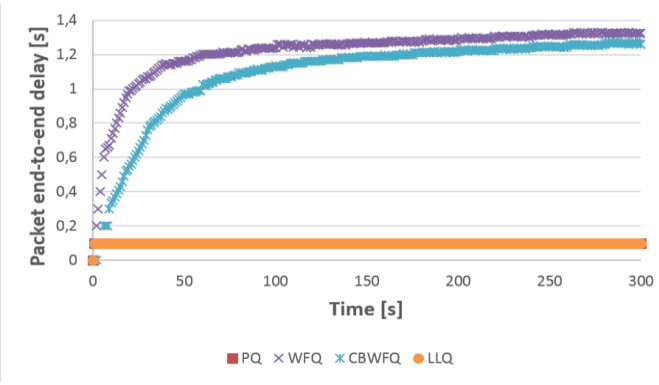


Fig. 7. VoIP traffic delay for PQ, WFQ, CBWFQ and LLQ

As for the delay variation (jitter) (see Fig. 8 and Fig. 9) the LLQ, PQ and CQ has managed to keep the measured values below 30ms. Subsequently, the WFQ and CBWFQ ensured around 150ms and 210 ms jitter, while the FIFO has finally managed to stabilize its delay variation around 1 second. It

should be noted that in terms of jitter PQ and LLQ congestion management algorithms managed to meet under the requirements of the QoS threshold requirement.

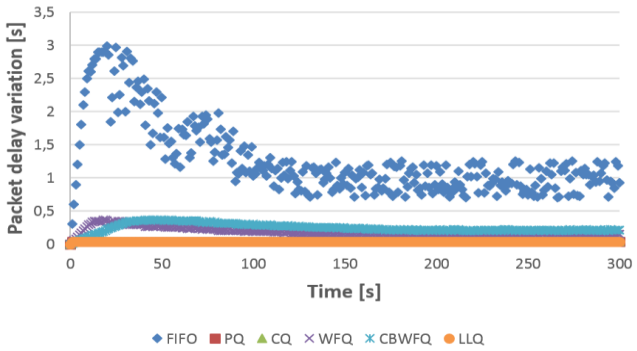


Fig. 8. VoIP jitter

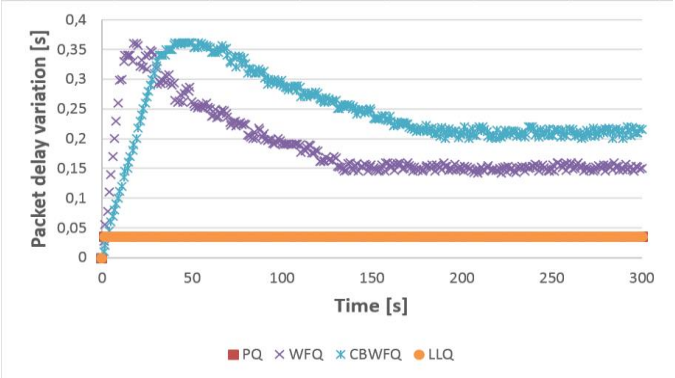


Fig. 9. VoIP jitter for PQ, WFQ, CBWFQ and LLQ

IV. CONCLUSION

This paper compares the performance of the main congestion management algorithms based on a measurement environment. The laboratory environment was implemented at the Faculty of Informatics University of Debrecen. In all cases the measurement result shows that the FIFO scheduling principle is the most inconvenient algorithm for handling of interactive voice packets in case of congestion. In the case of voice transmission the PQ and the LLQ algorithms were the two most appropriate algorithms, in terms of packet loss rate, end-to-end delay and jitter. Using these algorithms no packets suffered packet loss. However, knowing the principle of the PQ, namely that it serves the maximum priority queue, but produce packet starvation for other tree queues, based on the literature and on the measurement results the conclusion is, that for the interactive real-time voice traffic, taking all in consideration, the LLQ congestion management algorithm is most appropriate. Further research topic is to support the results and test the algorithms presented in the current article by mathematical modeling.

APPENDIX

These router configuration settings were used for implementing the congestion management algorithms:

FIFO

```
int s0/1/1
no fair-queue
end
```

PQ

```
access-list 101 permit tcp 172.16.10.0 0.0.0.255
172.16.20.0 0.0.0.255 eq 21
access-list 102 permit tcp 172.16.10.0 0.0.0.255
172.16.20.0 0.0.0.255 eq 1720
access-list 103 permit udp 172.16.10.0 0.0.0.255
172.16.20.0 0.0.0.255 eq 5060
priority-list 1 protocol ip high list 103
priority-list 1 protocol ip medium list 102
priority-list 1 protocol ip normal list 101
priority-list 1 default low
priority-list 1 queue-limit 20 40 60 80
int s0/1/1
priority-group 1
end
```

CQ

```
access-list 101 permit tcp 172.16.10.0 0.0.0.255
172.16.20.0 0.0.0.255 eq 21
access-list 102 permit tcp 172.16.10.0 0.0.0.255
172.16.20.0 0.0.0.255 eq 1720
access-list 103 permit udp 172.16.10.0 0.0.0.255
172.16.20.0 0.0.0.255 eq 5060
queue-list 1 protocol ip 2 list 103
queue-list 1 protocol ip 3 list 102
queue-list 1 protocol ip 4 list 101
queue-list 1 default 1
queue-list 1 queue 1 limit 4
queue-list 1 queue 2 limit 10
queue-list 1 queue 3 limit 10
queue-list 1 queue 4 limit 4
int s0/1/1
custom-queue-list 1
end
```

WFQ

```
int s0/1/1
fair-queue
end
```

CBWFQ

```
access-list 101 permit tcp 172.16.10.0 0.0.0.255
172.16.20.0 0.0.0.255 eq 21
access-list 102 permit tcp 172.16.10.0 0.0.0.255
172.16.20.0 0.0.0.255 eq 1720
access-list 103 permit udp 172.16.10.0 0.0.0.255
172.16.20.0 0.0.0.255 eq 5060
class-map VOICE
match access-group 103
exit
class-map VIDEO
match access-group 102
exit
class-map FTP
match access-group 101
exit
policy-map R1-Serial
class VOICE
bandwidth percent 30
exit
class VIDEO
bandwidth percent 30
exit
class FTP
bandwidth percent 10
exit
class class-default
bandwidth percent 5
exit
exit
int s0/1/1
no fair-queue
service-policy output R1-Serial
```

```
end
```

LLQ

```
access-list 101 permit tcp 172.16.10.0 0.0.0.255
172.16.20.0 0.0.0.255 eq 21
access-list 102 permit tcp 172.16.10.0 0.0.0.255
172.16.20.0 0.0.0.255 eq 1720
access-list 103 permit udp 172.16.10.0 0.0.0.255
172.16.20.0 0.0.0.255 eq 5060
class-map VOICE
match access-group 103
exit
class-map VIDEO
match access-group 102
exit
class-map FTP
match access-group 101
exit
policy-map R1-Serial
class VOICE
priority 384
exit
class VIDEO
bandwidth percent 30
exit
class FTP
bandwidth percent 10
exit
class class-default
bandwidth percent 5
exit
exit
int s0/1/1
no fair-queue
service-policy output R1-Serial
end
```

REFERENCES

[1] QOS, “Implementing Cisco Quality of Service, Student Guide”, Ver. 2.2, vol. 2, Cisco Systems Inc., 2006.

[2] A. S. Tanenbaum and D. J. Wetherall, “Computer Networks”, 5th Edition, Pretince Hall, ISBN-13: 978-0132126953, 2013.

[3] T. Svensson and A. Popescu, “Development of laboratory exercices based on OPNET Modeler”, Blekinge Institute of Technology, Department of Telecommunications and Signal Processing, 2003.

[4] C. Mancas and M. Mocanu, "QoS Optimization in Congested Multimedia Networks," Proceedings of the 36th TSP Conference, pp. 38-42, 2013.

[5] “Cisco IOS Quality of Service Solutions Configuration Guide”, Release 12.4T, Cisco Systems Inc., 2008.

[6] W. Odom, J. Geier and N. Mehta, “CCIE Routing and Switching Official Exam Certification Guide”, 2nd Edition, Cisco Press, 2006.

[7] A. S. Ranjbar, „CCNP ONT Official Exam Certification Guide”, Cisco Press, 2007.

[8] M. M. G. Rashed and M. Kabir, "A Comparative Study off Different Queuing Techniques in VoIP, Video Conferencing and FTP," Daffodil International University Journal of Science and Technology, vol. 5, no. 1, 2010.

[9] S. Szilágyi, "The Effects of Different Queuing Techniques over VoIP Performance: A Simulation Approach," in Abstracts & Pre-Proceedings of the 9th International Conference on Applied Mathematics, Baia Mare, Romania, 2013.

[10] S. Szilágyi, "Analysis of the Algorithms for Congestion Management in Computer Networks," Carpathian Journal of Electronic and Computer Engineering, vol. 6, no. 1, pp. 3-7, 2013.

[11] S. Akhtar, E. Ahmed, A. k. Saka and K. S. Arefin, "Performance Analysis of Integrated Service over Differentiated Service for Next Generation Internet," JCIT, vol. 1, no. 1, 2010.

[12] S. Szilágyi and B. Almási, "A Review of Congestion Management Algorithms on Cisco Routers," Journal of Computer Science and Control Systems, vol. 5, no. 1, 2012.

[13] A. Kuki, T. Bérczes, B. Almási, J. Sztrik, “A Queing Model to Study the Effect of Network Service Breakdown in a CogInfoCom System”, Cognitive Infocommunication (CogInfoCom) 2013, 4th IEEE International Conference on Cognitive Infocommunications, IEEE Publisher, pp. 205-2010, ISBN: 978-1-4799-1543-9, 2013.

[14] B. Almási, G. Bolch, D. Tutsch, “Stochastic Modeling of Multistage Interconnection Networks with MOSEL”, Journal of Mathematical Sciences, vol. 121, no. 5, 2004.

[15] L. L. Peterson and B. S. Davie, “Computer Networks: A System Approach - Network Simulation Experiments Manual”, 3rd Edition, Elsevier Inc., 2011.

[16] A. S. Sethi and V. Y. Hnatyshin, “The Practical User Guide for Computer Network Simulation”, NW: CRC Press Taylor & Francis Group, 2012.

[17] A. Cisco Networking, „Pagent IOS Tutorial,” Cisco Systems Inc., 2009.

[18] A. Cisco, "CCNP: Optimizing Converged Networks v5.0 - Instructor Lab Manual," Cisco Systems, Inc., 2007.

[19] C. Cisco, „TGN (Traffic GeNerator) User Manual,” Cisco Systems Inc., 2007.

Study of Gamification Effectiveness in Online e-Learning Systems

Ilya V. Osipov
i2istudy SIA
Krišjāņa Barona Iela, 130 k-10, Rīga, Lv-1012, Latvija

Evgeny Nikulchev
Moscow Technological Institute, ul. Kedrova, 8/2,
Moscow, Russia, 117292

Alex A. Volinsky
Department of Mechanical Engineering, University of South Florida, 4202 E. Fowler Ave., ENB118, Tampa FL 33620, USA

Anna Y. Prasikova
i2istudy SIA
Krišjāņa Barona Iela, 130 k-10, Rīga, Lv-1012, Latvija

***Abstract*—Online distance e-learning systems allow introducing innovative methods in pedagogy, along with studying their effectiveness. Assessing the system effectiveness is based on analyzing the log files to track the studying time, the number of connections, and earned game bonus points. This study is based on an example of the online application for practical foreign language speaking skills training between random users, which select the role of a teacher or a student on their own. The main features of the developed system include pre-defined synchronized teaching and learning materials displayed for both participants, along with user motivation by means of gamification. The actual percentage of successful connects between specifically unmotivated and unfamiliar with each other users was measured. The obtained result can be used for gauging the developed system success and the proposed teaching methodology in general.**

Keywords*—elearning; gamification; marketing; monetization; viral marketing; virality

I. Introduction

The paper describes newly developed open educational resource for learning foreign languages from native speakers, called i2istudy. Learning is achieved by using pre-defined educational materials through live interaction between the users acting as the teacher and the student. This is why the system is called i2istudy, "eye to eye", utilizing the peer-to-peer principle, based on the patented technology [1]. This technology allows learning basics of a foreign language from scratch, or enhance foreign language proficiency in a short period of time [2].

Open educational resources (OER) have recently become quite popular in the area of computer assisted language learning [3]. Currently there are several educational services on the market with a considerable amount of OERs that provide an opportunity to learn foreign languages (livemocha.com, www.learn-english-online.org, www.duolingo.com). [4, 5, 6] Most of these systems are automated, i.e. don't provide live human interactions. These systems can be divided into two categories: autonomous and social. Autonomous methods offer tasks, which are checked or monitored in accordance with the algorithms set up within the system as tests and quizzes, etc. Social methods allow direct or indirect interaction with real people, including communication and checking assignments, etc. (www.facebook.com). Such systems have been used in language learning [7, 8, 9, 10] also attempted to integrate computer-assisted language learning systems into the educational process.

Communication culture is formed as a result of live human speech interactions. Speech interaction is characterized by audio messages exchange between humans. Speech activity consists of the two aspects: language and speech. Together, they transform into the four types of speech activity, combined into two groups:

Receptive, perception-oriented types of speech activities, such as reading and listening;

Productive types of speech activities, focused on information production, such as speaking and writing.

Verbal means of communication are formed by all kinds of speech activity, which can be further developed and applied during foreign language learning. Thus, communication and training with representatives of other cultures is essential. An important quality for voice communication interaction is the "social" character of the learner, who's open for the dialogue and ready to participate in various discussions and debates. Learning a foreign language is also associated with acquiring the knowledge of other cultures, which is impossible without speech communication and knowledge of the linguistic and cultural features. Necessary properties of applications, which can provide language communication practice, are:

- The possibility of audio or visual contact, chosen by the training participants;
- Teaching methods, including conversation scenarios, allowing participants to actually start a conversation and keep it within the specified time on a given topic;
- Motivation of the participants.

Skype is currently the most common and popular tool for distance learning of foreign languages. While Skype was not

specifically designed for this purpose, it provides live audio and video connection between the participants. Thousands of small agencies and individuals offer distance language learning through Skype. The query for "English via Skype" in Google.com gives over 43 million results, with similar numbers when searched in Spanish, Russian and other languages. However, Skype does not allow finding people willing to teach or learn languages. It does not provide teaching or study materials, and does not track the time spent teaching or learning foreign languages. [11]

However, Skype and other systems of cooperative joint learning provide invaluable engagement, which plays a crucial role in learning. As noted by Clark and Mayer [12]: "all learning requires engagement", regardless of the delivery media. Zhang et al., [13] also suggested that increased student engagement can improve learning outcomes, such as problem solving and critical thinking skills. In the review article, Fredricks, Blumenfeld and Paris [14], defined engagement by its multifaceted nature: "Behavioural engagement draws on the idea of participation; it includes involvement in academic and social or extracurricular activities. Emotional engagement encompasses positive and negative reactions to teachers, classmates, academics, and school and is presumed to create ties to an object and influence willingness to do the work. Finally, cognitive engagement draws on the idea of mental investment; it incorporates thoughtfulness and willingness to exert the effort necessary to comprehend complex ideas and master difficult skills."

Fredricks, Blumenfeld and Paris [14], also claimed that the focus on behavior, emotion, and cognition, within the concept of engagement, may provide a richer characterization of learning. The authors reminded that a robust body of research addresses each of the components separately, but pointed out these dimensions of engagement had not been studied in conjunction. Thus, emotions aid communication process substantially [15,16].

Computer training system was developed, called i2istudy, in the form of a game to implement all three necessary components of spoken communication for training foreign language skills. This game consists of the computer-aided casual conversation with native speakers. For example, English-speaking users can learn Spanish from Spanish-speaking users, and visa versa. The i2istudy allows native speakers to teach others without knowing how to teach and without knowing foreign languages. In other words, i2istudy allows all native speakers, not necessarily professional teachers, to teach their native language in a collegial network game setting [17,18].

The main feature of the system consists of providing a common space with educational materials, including specifically designed lesson plans, which are simple and understandable step-by-step educational materials aiding communication. The platform, which allows live audio-video communication, is built into the web interface, based on the popular Web real time communications (WebRTC) technology.

Motivation is achieved by attracting a large number of users available online at the same time, always allowing to find a companion, along with gamification. Gamification is based on utilizing game elements in design and motivation principles in non-game situations [19]. It this case it is necessary to stimulate users to spend more time in the system to achieve the needed quantity and volume of practical skills, based on the modern principles realized in the e-learning systems [20]. The users should also be motivated to return to the system on the regular basis, the so called retention cycle [21].

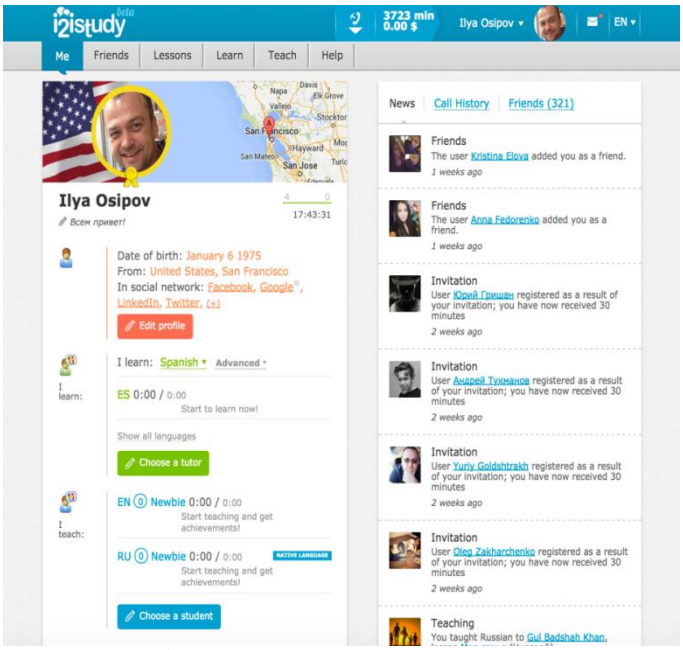


Fig. 1. User profile showing user achievement, how much time was used teaching and learning, along with other personal information

The developed application has the following gamification methods:

1) Time banking. When user acts as a student by taking lessons, virtual system currency in minutes is spent from the user account. One minute of learning is debited from the account, while one minute of teaching is credited to the account. Thus, the user acting as a teacher earns minutes, and the same user spends minutes as a student. In this way all users participate in the virtual economy. Users are motivated to earn minutes, pushing the user to periodically assume the role of a teacher. Each user gets 30 minutes in the system as a part of the registration process. If all minutes are spent in the account, the system does not allow to study, but offers to teach to earn more minutes. (Accumulated minutes are shown in figure 1) The implemented time banking goal is to motivate users to teach in addition to learning [22, 23].

2) Sequential lessons presentation. Most computer games utilize this gamification principle when the next game level becomes available after previous level has been completed.

For example, Figure 3 shows all lessons, but only a limited number of them is available. New lessons become available as the user goes through the previous lessons. Moreover, there is a grade displayed for each passed lesson as a single, dual or triple star, reflecting how well the student passed the test at the end of each lesson. Sequential opening of the lessons in batches intrigues the user to find out what's coming next, and boosts user engagement. Besides, explicit visibility of the grade encourages user to retake lessons with poor grades.

3) Achievements and badges. Figures 1 and 2 show user "achievements and badges". The user acquires nominal status, presented as an achievement, for learning and teaching in the system. The user gets status notifications by email, while other users also see these "achievements and badges", and can select their learning partner based on this information. Basic list of "achievements and badges" includes: "The First-grader; The Middle-schooler; High-school student". For short these are presented by the first two letters of the achievement, displayed in the corresponding language next to the user name, and are called badges. Shortened badges are used to save the space in the list, and will be replaced with medals in the future for better visibility. The goal here is to motivate users to receive awards as an external evaluation, thus motivating users to come back to the system and spend more time there.

4) Peer evaluation. For positive behavior reinforcement and polite communication between the users there is implemented peer evaluation. After each lesson both teacher and student can evaluate each other. There are two types of this kind of evaluation. The fist is simple like/dislike, which are accumulated for each user and displayed in the personal profile. This information is also visible to other users in the lists of teachers and students. Thus, polite and positive users are clearly visible, based on the large number of likes, while impolite and unpleasant users are also apparent due to dominating dislikes (Figures 1 and 2). Additionally, there is an option to report indecent user behavior to the system moderator. However, this option is a part of system moderation, rather than gamification.

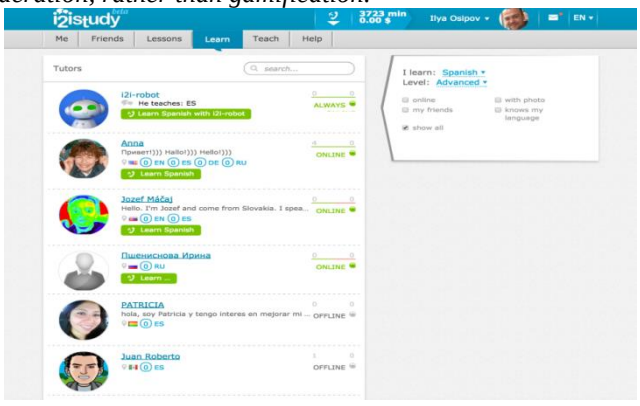


Fig. 2. Selecting a teacher. The list of users currently available as teachers. Every user can be called by pressing the green "Learn" button. If the "teacher" accepts the call, a lesson starts and lesson dialogue opens with live audio-video feed, shown in Figure 4

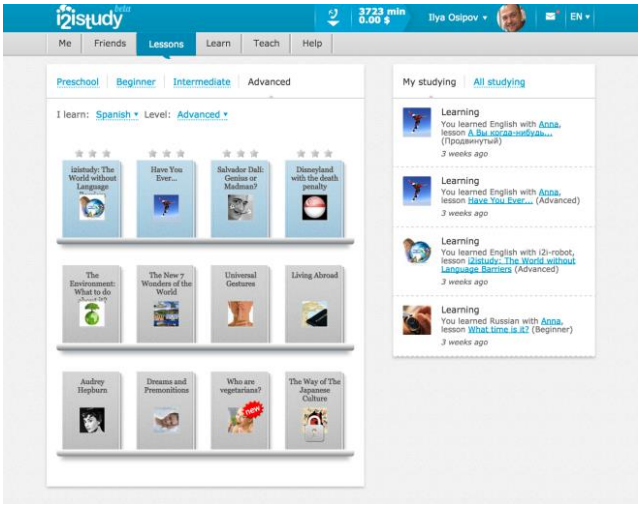


Fig. 3. The list of lessons. The blue lessons are available to the user, while the grey ones are not. Grey lessons become available as the blue lessons are passed, triggering user's interest and curiosity as a part of gamification

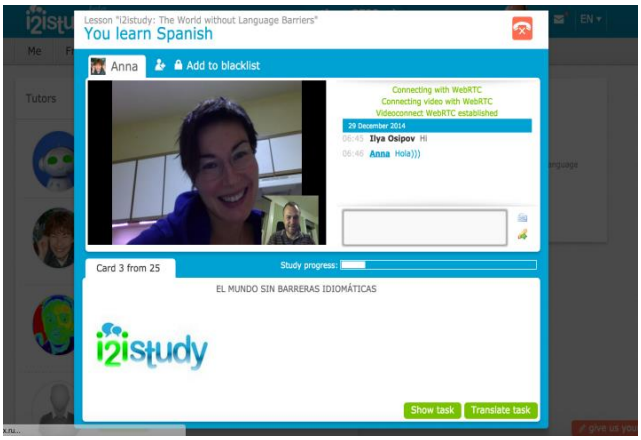


Fig. 4. Interface of a lesson in progress, where both users see and hear each other, along with the chat, study cards and the progress bar

II. Research Results and Discussion

It is necessary to evaluate the effectiveness of the mechanisms implemented in the application designed to improve verbal communication skills. Assessment is based on measuring:

- The increase in the number of application users, allowing to estimate the demand for service and implemented principles;
- The number of users willing to "teach", allowing to estimate the required number of people willing to teach and indirectly assess the suitability of the developed scenarios and gamification means of motivation;
- The time spent in the application as a parameter to a large extent characterizing the main development goal - the ability of users to establish long-term audio-video communication to obtain practical foreign language communication skills. The main research objective was to determine whether unfamiliar with each other

people, one acting as the "teacher", and the other acting as the "student", can take lessons together following a given scenario provided by the platform.

All experiments were conducted in the system with 40,000 registered users and 1,000-1,500 daily active users (DAU).

One of the main assumptions was whether unfamiliar with each other people, who met in the system for the first time, could communicate and to learn foreign languages together. The research objective was to find out if specifically unmotivated individuals without special training could choose a "teacher" or a "student" among the users currently available online in the system, send an invitation to study or to learn, establish audio-video connection and follow provided lesson scenario using the WebRTC face-to-face communication.

Initial system users, interested in practicing foreign language skills, registered in the system as a result of advertising in the Facebook social network. The ad suggested registering online and learning foreign languages for free, in exchange for teaching native language. This ad was displayed in Spanish and English-speaking countries, Germany and Russia. Besides, some users registered as a result of existing user invitations (users invited by the existing users). There was no additional information provided about the system, no verbal commentaries, or explanations were provided to the participants.

As a result, 39,729 users registered in the system in 6 months. 28,180 users indicated that they want to learn English, 8,711 Spanish, 1,028 Russian and 1,791 German languages. Wherein 14,943 users selected English as the native language, 20,673 Russian, 204 German and 3,843 Spanish. Monthly user registration data are shown in Table 1

TABLE I. NATIVE LANGUAGE OF THE NEWLY REGISTERED USERS PER MONTH

New users by languages

month	01.02 → 28.02	01.03 → 31.03	01.04 → 30.04	01.05 → 31.05	01.06 → 30.06	01.07 → 31.07	01.08 → 31.08	01.09 → 30.09	01.10 → 31.10	01.11 → 30.11	01.12 → 31.12	01.01 → 31.01	01.02 → 28.02	01.03 → 31.03	01.04 → 30.04	01.05 → 31.05	01.06 → 30.06	01.07 → 31.07	01.08 → 31.08
Russian language, users	5	4		5	296	515	1111	617	263	564	1069	3671	191	94	746	2186	3220	4329	1762
English language, users	3	1		1	44	87	171	99	32	80	178	746	68	25	295	978	4005	5582	2539
German language, users			1		11	9	15	4	5	6	15	22	3	3	6	22	29	38	14
Spanish language, users							1	1	3	1	17	43	12	3	79	760	903	1644	372

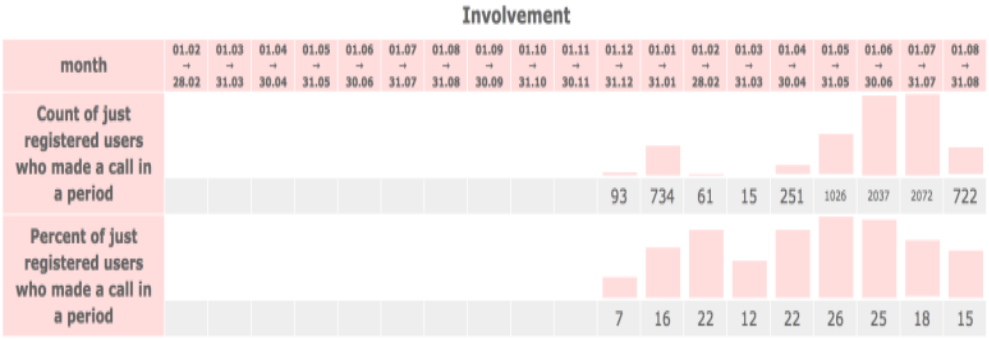
After registration users were asked to select a role of a teacher or a student and locate a potential learning partner to engage in the dialogue, based on the step-by-step methodology presented by the system. Each phrase is presented to the student with the corresponding prompts and translation if needed, and to the teacher with corresponding comments to aid the teaching process. The users see and hear each other in real time, while working with the synchronized teaching materials, and can type messages to each other in chat [24, 25].

Users could see other users in the system, along with their attributes, such as native language, the country of residence, the role of a "teacher" or a "student" in the system, along with the button to initiate a joint study session.

If a "teacher" accepts "student's" request, or visa versa, audio-video connection is established, where both participants can see and hear each other, along with the common filed with the synchronized study materials with the corresponding prompts for the teacher and the student. Besides, the system tracks the connection time for billing purposes in game currency, as seen in Figure 4.

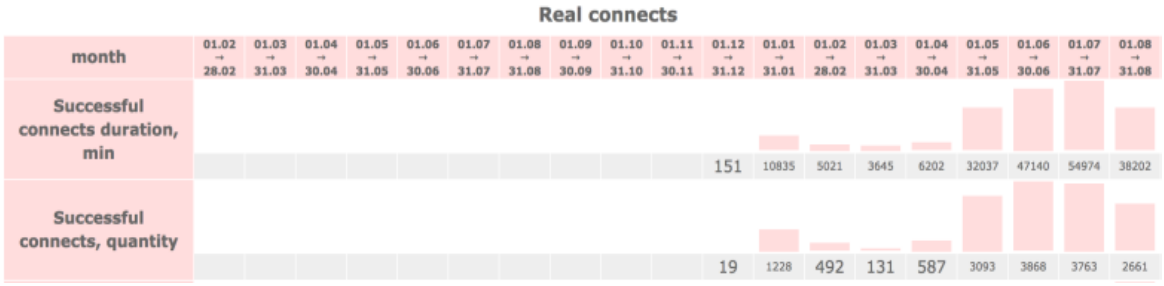
About 20% of all 40,000 registered users participated in the experiment. The rest were shy to speak with strangers, or decided not to spend their time. Some users failed to configure their microphone and the web camera needed for the real time audio-video connection, or their browser did not support Web RTC [26]

TABLE II. MONTHLY USER INVOLVEMENT



As a result of the conducted experiments it was established that two users, previously unfamiliar with each other, and met in the developed application for the first time could carry on a conversation following the suggested scenario, helping each other to learn foreign languages. Moreover, some users did not have a common language to use for communication. Average connection time was 11.94 min (189,207 min or 3,153 hours), divided by the number of successful connections (15,842). Any kind of interaction interruption was taken into account, including closing the browser or turning off the computer, or successfully finishing the lesson materials. Table 3 shows the number of successful connections and the connection duration in min.

TABLE III. THE NUMBER OF SUCCESSFUL CONNECTIONS AND THE CONNECTION DURATION



Regardless of the fact that the average connection time is not very long, the experiment showed that two unfamiliar and unprepared users can carry on a conversation in a foreign language for quite long. Besides, the average connection time continued to increase with the number of registered users, and reached 14.35 min in August 2014. Moreover, the most loyal and active users became apparent, spending hours learning and teaching, and even repeating the same lessons. Table 4 shows the most active users, along with the time spent learning or teaching in min.

TABLE IV. THE MOST ACTIVE USERS WITH TIME SPENT LEARNING OR TEACHING IN MIN

Successful tutor's connects duration, min			
duration	user id	name	Email
352	19131	L. M.	l.....@mail.ru
298	18418	j. b.	d.....@outlook.fr
277	15433	Л. Д.	h.....@mail.ru
276	18573	H. A.	n…..@mail.ru
260	28516	j. s.	m.....@gmail.com
254	391	A.	p.....@yandex.ru
216	22144	S. R.	s.....@gmail.com
213	20378	R. K.	m.....@gmail.com
200	1552	ج.ب	p.....@yahoo.com
145	29776	A. S.	a.....@yahoo.com
137	25718	Z. R.	z.....@hotmail.com
130	17253	R. M.	m.....@hotmail.com
126	26034	L.	s.....@yahoo.com
122	41271	A.	a.....@mail.ru
122	20179	D. V.	v.....@mail.ru
105	40693	E.	r.....@mail.ru
98	25965	A. M.	a.....@yahoo.com
98	40252	f. s.	f.....@yahoo.com
95	457	A. V.	b.....@yahoo.com
94	17364	A. H	a.....@gmail.com

The users registered as a result of advertising and conducted lessons either as a teacher or a student, learned the system interface on their own, without any special training. There were users were not specifically recruited to conduct initial proof of concept experiments. The users accepted roles or teacher and student. The numbers of both types of user roles are listed in Table 5. The corresponding ratio of 6.4 "teacher" users to 10.6 "student" users indicates that an average user is not afraid to play the role of a teacher.

TABLE V. THE NUMBER OF USERS WHO ACCEPTED THE ROLES OF "TEACHER" AND "STUDENT" FOR EVERY MONTH

	All new registered users	**New users who participated as a tutor**	**% tutors to all new**	**New users who participated as a student**	**% student to all new**
01.02 - 28.02	9				
01.03 - 31.03	5	0	0.0%	0	0.0%
01.04 - 30.04	1	0	0.0%	0	0.0%
01.05 - 31.05	6	0	0.0%	0	0.0%
01.06 - 30.06	362	0	0.0%	0	0.0%
01.07 - 31.07	782	0	0.0%	0	0.0%
01.08 - 31.08	1203	0	0.0%	0	0.0%
01.09 - 30.09	728	0	0.0%	0	0.0%
01.10 - 31.10	235	0	0.0%	0	0.0%
01.11 - 30.11	663	0	0.0%	0	0.0%
01.12 - 31.12	1646	19	1.2%	18	1.1%
01.01 - 31.01	4099	221	5.4%	370	9.0%
01.02 - 28.02	273	47	17.2%	97	35.5%
01.03 - 31.03	116	12	10.3%	22	19.0%
01.04 - 30.04	1504	172	11.4%	322	21.4%
01.05 - 31.05	4032	608	15.1%	993	24.6%
01.06 - 30.06	8319	620	7.5%	975	11.7%
01.07 - 31.07	11682	533	4.6%	965	8.3%
01.08 - 31.08	3969	287	7.2%	470	11.8%
Sum:	**39634**	**2519**	**6.4%**	**4232**	**10.7%**

Thus, the following experimental results were obtained:

The idea and the proposed form of training, based on the developed platform and teaching methods are in demand, indicated by the growth dynamics in terms of the number of users without motivational advertising.

Developed communication scenarios and lessons allows to overcome psychological barriers of communicating with strangers, both as a "student" and as a "teacher".

A significant percentage of users who want to act as teachers for the proposed method is revealed. This means that teaching staff is not required.

The introduced gamification adequately motivates users. A significant percentage of users have been returning to the application for further studies, demonstrating the effectiveness of developed tools and ideas.

III. CONCLUSIONS

Conducted study presents the new methodology to assess gamification tools in the e-learning systems. The e-learning system not only allows to conduct quality training, but presents and opportunity for statistical analysis of different parameters, contained in the log files, to assess the effectiveness of technical and pedagogical tools. Application popularity with users and grows of the number of users both act as assessment for the system motivation elements and tools.

Developed application is of interest for the majority of users, and allows to maintain a prolonged dialogue between the users in a given language. This definitely allows developing speech communication skills in a foreign language. Regardless of the stereotype that quality foreign language education can only be provided by the professional teacher, the developed system demonstrates that it is also convenient for users to study together. In this case, similar to the teaching materials presented in a text book, or interactive recorded media, professional teacher is recruited to develop teaching materials, while users can use these materials for training and practice. However, it is more interesting and encouraging doing this with other users, since the social effect also gets employed. Based on the conducted experiments, users not only spend more time in the system, but invite their friends to join them.

The authors suppose that similar ways of teaching could partially substitute individual tutoring and/or used as training to improve oral communication skills. It is concluded that the system should be developed further and recommended as the speech improvement tool. At the same time it is clear that the system is not a good fit for every user, since some people are very shy and cannot communicate with strangers, even when provided with pre-defined communication scenarios. Besides, the authors decided to reduce the average lesson duration to 15-20 minutes, as many users get tired, and only individual users continue communication for longer periods of time.

ACKNOWLEDGMENT

The authors would like to thank the i2istudy.com team members for their dedicated efforts: Vadim Grishin, Ilya Poletaev, Andrei Poltanov, Elena Bogdanova, Vildan Garifulin and Franziska Rinke.

REFERENCES

[1] Osipov I.V. (2014) SPONTANEOUS GROUPS LEARNING SYSTEM, US Patent 14546609 Filed November 18, 2014

[2] Benta, D., Bologa, G., Dzitac, I. (2014). E-learning platforms in higher education. Case study. Procedia Computer Science 31, 1170-1176.

[3] Coryell, J.E., Chlup, D.T. (2007). Implementing E-Learning components with adult English language learners: Vital factors and lessons learned. Computer Assisted Language Learning, 20(3), 263-278.

[4] Sevilla-Paóvn, A., Martínez-Sáez, A., Gimeno Sanz, A., Seiz-Ortiz, R. (2012). The role of social and collaborative networks in the development of in-house multimedia language learning materials. Procedia Social and Behavioral Sciences, 46, 1826.

[5] Giles, J. (2012). Learn a language, translate the web. New Scientist, 213, 18.

[6] Rutkin, A. (2014). The next wave of education. New Scientist, 222, 27.

[7] Donmus, V. (2010). The use of social networks in educational computer-game based foreign language learning. Procedia Social and Behavioral Sciences, 9, 1497.

[8] Toetenel, L. (2014). Social networking: a collaborative open educational resource. Computer Assisted Language Learning, 27(2), 149.

[9] Aydin, S. (2014). Foreign language learners' interactions with their teachers on Facebook. System, 42, 155.

[10] Tal, M., Yelenevskaya, M. (2012). Computer-assisted language learning: Challenges in teaching multilingual and multicultural student populations. Procedia Social and Behavioral Sciences, 47, 263.

[11] Rao, B., Angelov, B., Nov, O. (2006). Fusion of disruptive technologies: Lessons from the Skype case. European Management Journal, 24(2-3), 174.

[12] Clark, R. C., & Mayer, R. E. (2011) E-learning and the Science of Instruction. Pfeiffer (San Francisco).

[13] Zhang, D., Zhou, L. Briggs, R. O., & Nunamaker, J. F. (2006) Instructional video in e-learning: Assessing the impact of interactive video on learning effectiveness. Inform. manage. 43 (1), 15-27, doi: 10.1016/j.im.2005.01.004.

[14] Fredricks, J. A., Blumenfeld, P. C., & Paris, A. H. (2004) School Engagement: Potential of the Concept, State of the Evidence. Rev. educ. res. 74 (1), 59–109, doi:10.3102/00346543074001059.

[15] Lang, P. J. (1995) The emotion probe: studies of motivation and attention. Am. psychol. 50 (5), 372-385, doi : 10.1037/0003-066X.50.5.372.

[16] Martens, R. L., Gulikers, J., & Bastiaens, T. (2004) The impact of intrinsic motivation on e-learning in authentic computer tasks. J. comput. assist. lear. 20 (5), 368–376, doi: 10.1111/j.1365-2729.2004.00096.X.

[17] Buga, R. Cǎpeneață, I., Chirasnel, C., Popa, A. (2014). Facebook in foreign language teaching – A tool to improve communication competences. Procedia Social and Behavioral Sciences 128, 93-98.

[18] Zolfaghar, K., Aghaie, A. (2012). A syntactical approach for interpersonal trust prediction in social web applications: Combining contextual and structural data. Knowledge-Based Systems 26, 93-102.

[19] Domínguez, A., Saenz-de-Navarrete, J., De-Marcos, L., Fernández-Sanz, L., Pagés, C., & Martínez-Herráiz, J. J. (2013). Gamifying learning experiences: Practical implications and outcomes. Computers & Education, 63, 380-392.

[20] Bubnov G., Nikulchev, E., Pluzhnik, E. (2015) "Experience the effective implementation of innovative information technologies in educational institute", Vysshee obrazovanie v Rossii [Higher Education in Russia], 1, 159–161 (In Russian)

[21] Osipov I.V., Volinsky A.A., Grishin V.V. (2014) Gamification, virality and retention in educational online platform. Measurable indicators and market entry strategy. arXiv preprint arXiv:1412.5401

[22] Marks, M. (2012) Time banking service exchange systems: A review of the research and policy and practice implications in support of youth in transition. Children and Youth Services Review, 34(7), 1230-1236.

[23] Válek, L., Jašíková, V. (2013) Time bank and sustainability: The permaculture approach. Procedia Social and Behavioral Sciences, 92, 986-991.

[24] Hye Yeong Kima (2012) Learning opportunities in synchronous computer-mediated communication and face-to-face interaction. Computer Assisted Language Learning Volume 27, Issue 1, 2014

[25] Markus Kötter (2001) MOOrituri te salutant? Language Learning through MOO-Based Synchronous Exchanges between Learner Tandems. Computer Assisted Language Learning. Volume 14, Issue 3-4

[26] Osipov I.V. (2014) "Indicators of Viral and Retention for Freemium Product. Market Entry", Cloud of Science, 1(1), 457-471 (In Russian)

The Real-Time Research of Optimal Power Flow Calculation in Reduce Active Power Loss Aspects of Power Grid

Yuting Pan[1]
College of Electronic and Electrical Engineering
Shanghai University of Engineering Science
Songjiang District, Shanghai 201620, China

Yuchen Chen[2]
College of Electronic and Electrical Engineering
Shanghai University of Engineering Science
Songjiang District, Shanghai 201620, China

Zhiqiang Yuan[3]
Shanghai Electric Power Design Institute Co., Ltd.
Huangpu District, Shanghai 200025, China

Bo Liu[4]
Shanghai Electric Power Design Institute Co., Ltd.
Huangpu District, Shanghai 200025, China

***Abstract*—In order to research how to availably reduce the active power loss value in power grid system when the power system is operating, it offers a quantitative research in theory through conceiving the unbalanced losses of power grid system under the overloading bus as the investigative object, and establishing an active power loss mathematical model. It carries out online real-time optimal flow calculation within the condition that meets the control variables and state variables of the equality and inequality constraints. For some branches with larger network loss, it respectively adopts three methods, including voltage regulation method, reactive power compensation method, changing the branch's cross-sectional area method, to reduce the general active power loss values. Moreover, it compares the compensation equivalent of three methods during the recovery process of the general active power loss in the power grid. Taking IEEE14 as an example, it verifies the effectiveness of the proposed methods. It not only can offer a reasonable measure to reduce the losses of power grid, but can provide some reliable reference for the power grid dispatching personnel.**

Keywords—optimal power flow; voltage regulation; reactive power compensation; cross-sectional area; active power loss

I. INTRODUCTION

In recent years, with the increasement of electricity power c onsumption, the scale of the power grid has been enlarged year by year, the power loss in the electricity power system is becoming a more and more serious problem, the research of how to reduce the active power loss not only be comes the major factor of reducing the cost of electricity power system, but also becomes the research direction of the future's electrical power supply.

Wide Area Measurement System (WAMS) technology, which has a main characteristic of collecting data in distributed, has been widely concerned. The main principle of WAMS technology is: real-time acquisition of voltage, phase data and relative information from Phasor Measurement Unit (PMU) in different grid, comparing with a synchronously upload to the computer, therefore monitoring and controlling the overall loss of power system [1-4]. Calculating the active power loss by computer power flow to realize the real time measurement of power grid, there are generally 3 basic steps: ① Establishing mathematical model of the problem; ② Using an efficient analysis method; ③ The preparation of the relevant software program [5-8].

This article mainly studies on how to find the combined me asure to decrease the active power loss in the power grid. It sets up the mathematical model of the active power loss, and uses the independently developed online real-time optimal Newton-
Raphson power flow calculation software to calculate cases.

This article adopts three decreasing network loss methods, including voltage regulation method, reactive power compensation method, changing the branch's cross-sectional area method, to start power flow calculation for the IEEE 14-bus system and analyze the compensation equivalent of three methods. It can offer a reliable basis for power grid dispatching personnel to adopt a more effective way so as to decrease the power loss in the condition of the load overload.

II. A MATHEMATICAL MODEL OF POWER GRID OPTIMAL FLOW CALCULATION

A. The objective function of optimal power active power loss of power network

The intelligent vacuum circuit breaker online monitoring system is composed by the host computer, the lower computer and signal processing modules. The lower computer hardware platform consists of a TMS320F2812 DSP and peripheral hardware circuits. It collects mechanical parameters, divide-shut brake circuit current signal and vibration signal of vacuum circuit breaker. The host computer uses ARM as a core, mainly working as remote communication with the host computer, and getting the results of data processing and eigenvalues from the DSP at the same time. The preprocessed data is transferred from DSP to ARM via an HPI interface, and transmitted via the Ethernet interface to the host computer for data analysis

and processing, so as to determine the current status of the circuit breaker, and analyze its operation situation and diagnose malfunctions. The system structure is shown as Fig.1.

Through the analysis of regression loss mathematical model that can improve accuracy of network loss in the calculation of load variation which presented in the literature 6, in the case of consulting a lot of related information, successfully developing a kind of power network active power loss mathematical model which is suitable for online real-time power flow calculation, the objective function constructed as shown in equation (1):

$$\min\ f(x) = \min \sum_{i,j \in l} \left(\Delta P_{ji} + \Delta P_{ij} \right) \qquad (1)$$

In the formula, ΔP_{ji} is the active power loss value from the j node flow to the i node, ΔP_{ij} is the active power loss value from the i node flow to the j node.

B. The system constraints of optimal power flow [9]

The active power loss objective function (1) should satisfy the constraints shown in formula (2), the constraint condition in the common constraint is introduced in the line flow constraints, in order to reasonably control the reactive power load distribution.

$$\begin{cases} P_{Gi} - P_{Li} - e_i \sum_{j=1}^{n} (G_{ij} e_j - B_{ij} f_j) - f_j \sum_{j=1}^{n} (G_{ij} f_j - B_{ij} e_j) = 0 \\ Q_{Gi} - Q_{Li} - f_i \sum_{j=1}^{n} (G_{ij} e_j - B_{ij} f_j) - e_j \sum_{j=1}^{n} (G_{ij} f_j - B_{ij} e_j) = 0 \\ S_{ij}^2 = P_{ij}^2 + Q_{ij}^2 \le S_{ij\max}^2 \qquad i, j = 1, 2, \ldots, n \\ V_{i\min} \le V_i \le V_{i\max} \qquad i = 1, 2, \ldots, n \\ P_{Gi\min} \le P_{Gi} \le P_{Gi\max} \qquad i = 1, 2, \ldots, n_g \\ Q_{Gi\min} \le Q_{Gi} \le Q_{Gi\max} \qquad i = 1, 2, \ldots, n_g \end{cases} \qquad (2)$$

In the type, P_{Gi} and Q_{Gi} are the active power and reactive power output of generator i respectively; P_{Li} and Q_{Li} are the active power and reactive power load of generator i respectively; e_i and f_i are the real part and imaginary part of the node voltage of i respectively; V_i is the voltage value for node i; G_{ij} and B_{ij} are the conductance and the susceptance of the branch ij; n_g is the number of generator; S_{ij}, P_{ij} and Q_{ij} are the apparent power, active power and reactive power of the branch ij respectively.

III. A CASE OF OPTIMAL POWER FLOW CALCULATION IN ACTIVE POWER LOSS

The Fig.1 is a IEEE14 node grid, it consists of 20 branches, 14 nodes which contains 3 transformer branch and 17 transmission lines, in the 5 node of power plant, the 1# node is the balance node, the rest of the load nodes is the PQ node, the power plant node is the PV node。

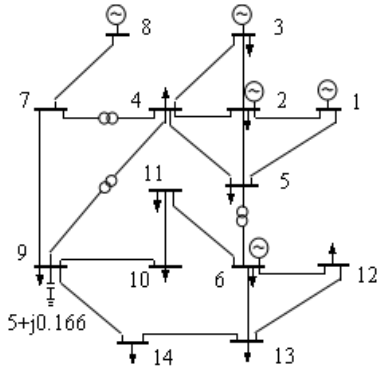


Fig. 1. IEEE14 typical grid topology

For verifying the loss compensation ability of the IEEE14 node power grid in the long distance transmission, this paper selects the No.9 bus which is the furthest node from all generators in the theory as the research object. Assuming the No.9 bus overloads, its load value is equal to 1.25 times the normal load conditions, that is, the load value is $0.36875 + j0.2075$ MW, all the remaining data of IEEE14 node voltage are consistent with the normal conditions. Through the power flow calculation, we can know the before and after change of the general active power loss values of power network, specific data is indicated in TABLE I.

TABLE I. THE UNCHANGING AND CHANGING LOAD VALUE $\Delta P'_{LOSS}$ OF TOTAL GRID ACTIVE POWER LOSS IN 9 BUS

Name	The load value of No. 9 bus/MW	The power loss $\Delta P'_{LOSS}$ /MW	The variation of network general active power loss increment value $\Delta P'_{LOSS}$/MW
Initial Value	0.295+j0.166	0.195477	0.008618
Change Value	0.36875+j0.2075	0.204095	

As it can be seen from the TABLE I, after the load of No.9 bus has increased 1.25 times than before, the general active power loss increment value $\Delta P'_{LOSS}$ of network is 0.008618MW.

By means of three kinds of methods, including voltage regulation, reactive power compensation, changing conductor cross-sectional area, it computes the general power active power loss variation in every methods by the online real-time power flow calculation software, then let them respectively compare with the value 0.008618. After that, it gets the corresponding comparative quantities $\Delta P_{LOSS}1$、$\Delta P_{LOSS}2$、$\Delta P_{LOSS}3$. The final analysis of compensation equivalent conditions among three kinds of methods will be given.

A. The regulation of No.2 bus voltage

From the flow computational results, the branch 1-2 has the largest network loss. As we know, the No.1 bus is the balance of nodes, therefore it considers to regulate the voltage of No. 2 bus so as to effectively reduce the network loss. According to the data of the power grid， it analyzes the function relationship between U_2 and $\Delta P_{LOSS}1$:

$$\Delta P_{LOSS}1 = 9.6381165 - 17.9458 * U_2 + 8.355 * U_2^2 \quad .$$

From Fig.2, it can be more intuitive to observe the corresponding changes.

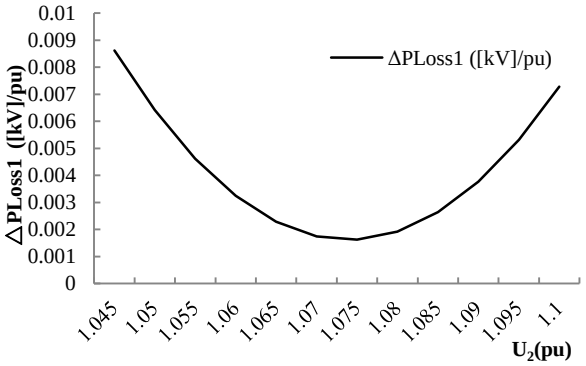


Fig. 2. The changing situation of $\Delta P_{LOSS}1$ with elevating of the value in U_2

It can be seen from the Fig.2, with the value U_2 of No.2 bus voltage is gradually increasing， the general active power loss comparison value $\Delta P_{LOSS}1$ of power grid shows a decreased first and then increased trend, it can be in accordance with the advantages of reducing loss to adjust U_2 value， in order to achieve the best loss reduction effect.

B. Reactive power compensation of No.4 and No.5 bus

From the power flow calculation, we can know that the reactive power near 4 and No. 5 bus is lower, therefore we should use compensatory of result on the spot scheme to compensate reactive power for No.4 and No.5 bus, so that it can improve active power loss value of the whole grid to achieve the purpose of reducing network loss. Specific data is shown in the TABLE Ⅱ.

TABLE II. THE VALUE OF $\mathrm{D}P_{LOSS}2$ WITH COMPENSATING $\mathrm{D}Q_C$ IN NO.4、NO.5 BUS

$\Delta Q_c/(pu)$	$\mathrm{D}P_{LOSS}2/([M\ \mathrm{var}]/pu)$
0	0.007990106
0.15	0.006042856
0.3	0.004570166
0.45	0.003456381
0.6	0.002614035
0.75	0.001976975
0.9	0.001495171
1.05	0.001130786

The TABLE Ⅱ shows the reactive power compensation increment $\mathrm{D}Q_C$. It is the reactive power compensation value of No.4 and No.5 bus is the value that is compared with 0.008618MW after compensating for corresponding. According to the calculated data, it gets the function relationship:

$$\mathrm{D}P_{LOSS}2 = 0.007990106 * e^{-1.862181813*VQ_C}.$$

The corresponding situation can be seen more intuitively from the Fig.3.

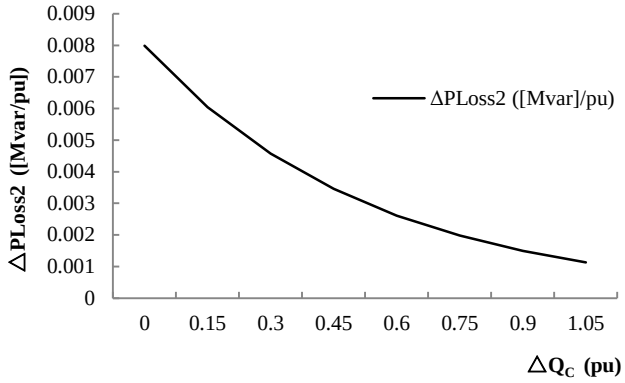


Fig. 3. The changing situation of $\mathrm{D}P_{LOSS}2$ with compensating $\mathrm{D}Q_C$ in NO.4、NO.5 bus

From the Fig.3, with the reactive power compensation increment $\mathrm{D}Q_C$ increasing gradually, the compared value of network general active power loss increment value $\mathrm{D}P_{LOSS}2$ has shown a trend of decreasing.

C. Changing the branch's cross-sectional area which the network loss is larger

By the power flow calculation, the IEEE14 nodes with grid network loss of the top six branches in descending order in turn are as follows in TABLE Ⅲ.

TABLE III. THE NETWORK LOSS DESCENDING ORDER OF IEEE14 NODE POWER GRID BRANCH

Initial node	Termination node	Active power loss	The ranking of net loss
1	2	0.090176	1
1	5	0.035943	2
2	3	0.026037	3
3	4	0.019457	4
2	4	0.012471	5
2	5	0.006363	6

Considering in this theoretical case, it changes the cross-sectional area of different branches respectively, The corresponding grid changes of the general active network loss comparison value is shown in Fig.4.

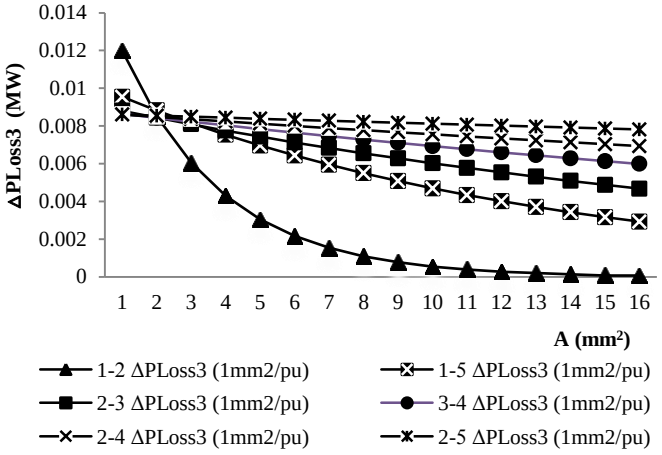


Fig. 4. The changing situation of $\Delta P_{LOSS}3$ with 1 mm^2 increment of cross-sectional area values ranked the top six in power grid loss

The Fig.4 shows that with the increasing of branch cross-sectional area value, and it all shows a trend of decreasing network loss. Among them, the loss reduction effect of branch 1-2 is most obvious, while the effect of branch 2-5 is the most unobvious one.

D. To compare compensation equivalent of the three methods Changing the branch's cross-sectional area which the network loss is larger

From the above analyses， we can conclude that each of three methods discussed can reduce the general active power loss in some degree. The effects of three methods in the view of loss reduction are compared in order to get compensation equivalent. The relationships among the three methods and the compared values of general active power loss are shown in Fig.5.

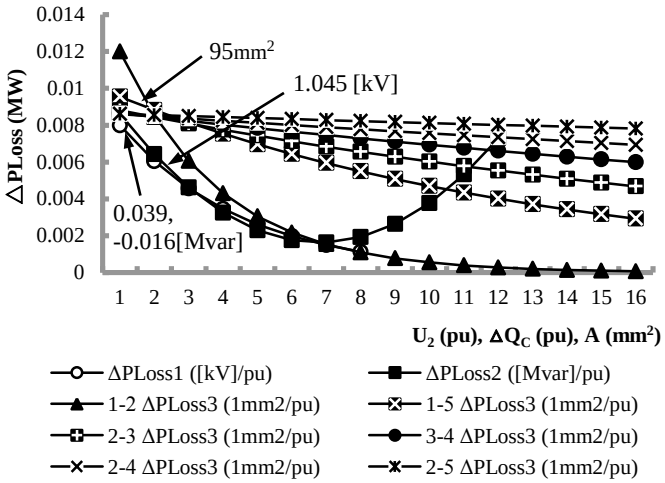


Fig. 5. The relationships among the three methods and the compared values of power grid loss

According to the Fig.5, the initial reactive power values of No.4 and No.5 bus are 0.039 [*Mvar*] and -0.016 [*Mvar*] in reactive power compensation method. The initial voltage of bus is 1.045 [*kV*] in voltage regulation method. The initial area value of each branch is 95 mm^2 in changing the branch cross-sectional area method.

In TABLE Ⅳ, it shows compensation equivalent of each method when ΔP_{LOSS} is reduced from 0.008MW to 0.006MW.

TABLE IV. THE NETWORK LOSS DESCENDING ORDER OF IEEE14 NODE POWER GRID BRANCH

Name / (unit)	Compensation equivalent
ΔU_2 /(*pu*)	0.005
ΔQ_C /(*pu*)	0.15
A(1- 2)/ (mm^2)	0.9

According to the principle that the increasing of the loss maximum branch cross-sectional area value of power grid can improve loss reduction effect better, the cross-sectional area compensation equivalent in the TABLE Ⅳ just considers changing the cross-sectional area value of the branch 1-2. In the Fig.5, it lists the changing relationship between the top six branches of the power grid loss and the general power grid active power loss comparison values. Because of the actual situation of cross sectional area of the increment value can't be 1 mm^2 , if we consider the actual factors, it changes the cross-sectional area value of branch 2-5 will be obtained with a more reasonable compensation equivalent than the two method of voltage regulation and reactive power compensation.

IV. CONCLUSIONS

This article uses the online software of optimal power flow calculation to calculate and analyze three methods' effect on the general active power loss of power network loss reduction in the condition of No.9 bus, which is the more distal bus to the generator in the network of IEEE 14 nodes, overloads. We can get three methods' compensation equivalent. The specific conclusions are as follows:

1) If the value of U_2 raises, the change trend of network general active power loss increment value $\Delta P_{LOSS}1$ is decreased first and then increased. When the value of U_2 is about 1.074 [kV], the effect of reducing the loss is the best.

2) The increasing of 0.15 unit in Q_C corresponds to the increasing of 0.005 unit in U_2, which means that the way of voltage regulation is better than reactive power compensation in the effect of reducing the loss.

3) When it changes the value of cross-sectional area in different branches, the general active power loss comparison value $\Delta P_{LOSS}3$ shows a monotonic decreasing trend. Also, it has a more obvious effect in reducing the loss if it raises cross-sectional area value of the branch which has a larger network loss.

4) Both reactive power compensation method and changing the branch's cross-sectional area method need to consider the additional expenses, so they are not economic in practical engineering application. Therefore, we should first consider it to drop network loss through voltage regulation

method. When the per unit value of U_2 is up to 1.074 [kV], then we can take changing the branch cross-sectional area method in order to realize the aim of decreasing the network general active power loss effectively.

Through the study of different methods of reducing the network general active power loss, we can provide reliable information for the dispatchers. This is helpful to analyze how to reduce the actual power grid loss effectively. In the future, the research direction will be dedicated to the combination of synchronous phasor measurement unit and the present research, to realize the purpose of on-line measurement of the network loss among different power grid, thus it can provide more useful information for the power grid and relevant departments.

ACKNOWLEDGMENT

The project has been supported by Chinese National Natural Science Foundation (No.51177099), Shanghai City Committee of science and technology project （No.10160501700).

REFERENCES

[1] XI Peiqi, WU Miaofeng. Application of Synchronized Phasor Measurement Device in Far East 500 Kv Substation[J]. East China Electric Power, 2014, 42(7) : 1480-1482.

[2] CHENG Yunfeng, ZHANG Xinran, LU Chao. Research progress of the application of wide area measurement technology in power system[J]. Power System Protection and Control, 2014, 42(4) : 145-153.

[3] DONG Qing, ZHAO Yuan, LIU Zhigang, et al. A Locating Method of Earth Faults in Large-scale Power Grid by Using Wide Area Measurement System[J]. Proceedings of the CSEE, 2013, 33(31) : 140-146, S17.

[4] WU Xing, LIU Tianqi, LI Xingyuan, et al. Optimal Configuration of PMU Based on Data Compatibility of WAMS/SCADA and Improved FCM Clustering Algorithm[J]. Power System Technology, 2014, 38(3) : 756-761.

[5] SUN Qiuye, CHEN Huimin, YANG Jianong, et al. Analysis on Convergence of Newton-like Power Flow Algorithm[J]. 2014,34(13) : 2196-2200.

[6] AN Sicheng, WU Kehe, BI Tianshu, et al. Synchronization Algorithm for Real-Time Data Concurrent Access Applicale to WAMS[J]. Proceedings of the CSEE, 2014, 34(19) : 3226-3233.

[7] XU Yan, LU Bin, WANG Zengping. A Power Flow Transfer Identification Scheme Baesd on WAMS [J]. Proceedings of the CSEE, 2013, 33(28) : 154-160, S21.

[8] ZHAO Yuanyuan, CUI Yong, AI Qian. Dynamic Adjustment Schemes for Divisional Power Grid Structures Based on Phasor Measurement Unit[J], East China Electric Power, 2014, 47(5) : 0859-0864.

[9] CHEN Liang, BI Tianshu, LI Jinsong, et al. Dynamic State Estimator for Synchronous Machines Based on Cubature Kalman Filter[J]. Proceedings of the CSEE, 2014, 34(16) : 2706-2713.

Assessment of Potential Dam Sites in the Kabul River Basin Using GIS

RASOOLI Ahmadullah
Department of Information Engineering
University of the Ryukyus
Okinawa, Japan

KANG Dongshik
Department of Information Engineering
University of the Ryukyus
Okinawa, Japan

***Abstract*—The research focuses on Kabul River Basin (KRB) water resources infrastructure, management and development as there are many dams already in the basin and many dams are planned and are being studied with multi-purposes objectives such as power generation, irrigation and providing water to industry and domestics.**

KB has been centralized all water resources related information in an integrated relational geo-database this KB is centralized repository for information river basin management with the main objectives of optimizing information collection, retrieval and organization. In addition, in this paper information and characteristics of the KRB has been presented such as drainage network or hydrology, irrigation, population, climate and surface pattern other necessary features of the basin by the use of GIS in order to invest and implement infrastructure projects.

The first step in doing any kind of hydrologic modeling involves delineating streams and watersheds, and getting some basic watershed properties such as area, slope, flow length, stream network density, etc. Traditionally this was (and still is) being done manually by using topographic/contour maps. With the availability of Digital Elevation Models (DEM) and GIS tools, watershed properties can be extracted by using automated procedures.

The processing of DEM to delineate watersheds is referred to as terrain pre-processing. Besides that, it produced the necessary thematic maps, base maps and other detailed maps for illustrating basin characteristics and features GIS Based.

Keywords*—*Geographical Information System (GIS); Kabul River Basin (KRB); Digital Elevation Model (DEM); Map

I. INTRODUCTION

Geographical information system (GIS) is an efficient tool for analyzing, collecting, storing, manipulating, displaying, editing vector and raster data for particular purposes. So using GIS can also play a crucial role in assessing potential sites for dam selection by hydrological analysis and modeling.

Kabul RB has 35% population density of the country and Fifty-nine percent population of the basin is rural and lives outside Kabul; more than 96 percent live in small villages and settlements, primarily along the rivers in cultivable areas with Access to water. Rain-fed agriculture is only approximately 3 percent of the total cultivated area in the basin.

Agriculture constitutes is the major income source for the population in the Kabul Basin study area and the economic development is in many respects connected to the presence of water resources and their rational use.

Over the last 40 years, there have been very intensive human-induced environmental changes in this area, primarily associated with irrigation activities, changes in the grazing pressure on desert rangelands and deforestation.

In addition many dams have at least some flood mitigation effects to their primary purposes. However for the assessment of potential dam sites in the KRB created Knowledge Base (KB) and Geo-database also produced thematic maps, base maps and required detail maps for the basin water resources infrastructure development projects as well as water resources management in the basin.

Dams are one of the prime options to store and use the water more efficiently and improve living condition for rural and urban population by providing electricity, water for industry, irrigation and drinking [1].

In this paper KB has been centralized all water resources related information in an integrated relational geo-database this KB is centralized repository for information river basin management with the main objectives of optimizing information collection, retrieval and organization. So besides information and characteristics, KRB has also been explained such as drainage network/system or hydrological features, irrigation, population, climate and so on illustrated as maps and graphs.

The presented methodology and required data sources can be the basis for an actual pilot study on a limited area to test and to calibrate. Thus for Hydrological analysis and modeling Areal-photos, Satellite image and DEM 90m SRTM also vector data has been processed, digitized and also extracted and computerized all required surface features data and information stored as database.

II. STUDY SITE

A. The principle sub-basins of Kabul River Basin

Kabul RB is a 700 kilometer long river that starts in the Sanglakh Range of the Hindu Kush Mountains in Afghanistan and ends in the Indus River near Attock, Pakistan [2]. This RB has been located in south-east of Afghanistan and it is the portion of the Indus River catchments. the total catchment area of Kabul RB is about 76,908 (Sq km) including 14,000 (sq km)

of the upstreams sub catchments which is located in Pakistan. And many big provinces and cities are situated in the basin such as Wardak, Kabul, and Jalalabad so on, and also major tributaries are considered in the Basin like Kunar, Logar, Kabul and Panjsher.

The eastern part of the basin which is originated from Pakistan has higher elevations, and covered by snow most of the year. About 72% of total runoff is originated from this part of the basin. Because of the elevation variation (400-6,000 m) As shown in Fig. 1, This part of basin has considerable potential to install hydropower.

Western part of the basin is relatively dry shortage for water especially in dry season [3].

From the standpoint of climate, hydrology, and physiographic characteristics, the Kabul River basin is divided into three distinct sub basins as shown in Fig. 1.

The upper basin consists of two major sub basins the Panjshir sub basin and the Logar-Upper Kabul sub basin and. The third sub basin is the Lower Kabul, which encompasses the watershed area from the confluence of the Panjshir and Upper Kabul rivers near the head of the Naglu reservoir to the border with Pakistan and Shmal Khuram watershed includes mostly Paktia and Khost provinces.

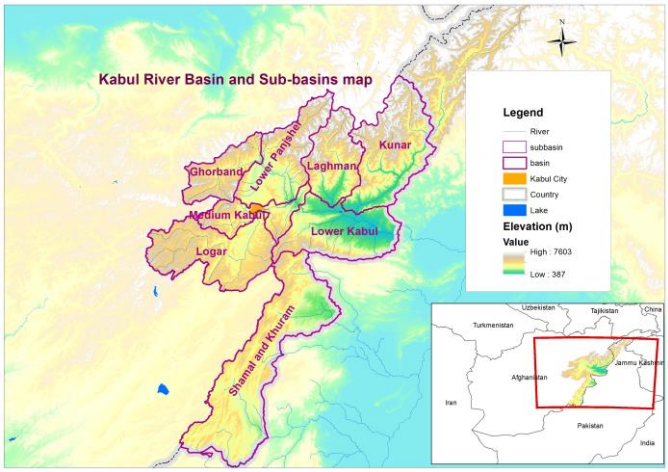


Fig. 1. Shows Kabul River Basin's and Sub-basins

B. The basin designed Maps

These maps have been produced in three sections with various scale such as (1:2000, 000, 1:500,000 and 1:750,000 etc).

a) Base Maps such as (river and sub-basins base map, river infrastructure and irrigation, hydro-meteorological stations, administrative units, basin administrative and management, population and population concentration, topography/elevation, regional setting (Indus Basin)).

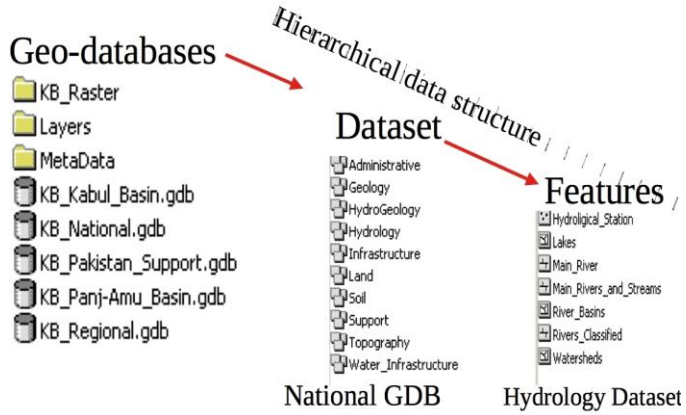
b) Thematic Maps have also been made for the basin namely precipitation and temperature regime, hydrographs, snow coverage, energy infrastructure, soil taxonomy, land use and land cover, terrestrial ecoregions, geology, geological faults respectively.

c) The more detailed sub-basins Maps consisted of sub-basins lower Kabul and Kunar, sub-basins Middle Kabul and Panjsher and sub-basins Logar/ Shamal and Khuram respectively.

Therefore each section has consisted of different maps with multi-purposes created by using GIS and Arc-hydro tools and it's applications for water resources development and infrastructure projects.

III. Knowledge Base and Geo-Database

Geo-database and Knowledge Base in Arc GIS are organized and contain of the required data (raster and vector) As national level for Kabul basin. Fig. 2, shows the hierarchical data structure of the geo-database formed of



datasets and feature classes.

Fig. 2. Shown the structure of geo-database in Arc GIS

IV. Data Used

By using Arc GIS tools based on required data for the basin, Arc-hydro and it is applications hydrological analysis, mapping (thematic, base) and modeling, raster data such as (Areal-photos, Satellite images and Digital Elevation Model (DEM) (90m SRTM) Landsat imagery) and also vector as well as statistic and historical series data from Central Statistics Office (CSO) has been used and processed, also including information on the hydrological network, irrigation, energy infrastructure and general themes such as climate, soils, geology, land use and ecosystems etc.

The raw data and information related to the study area Kabul basin were obtained from various sources like Afghanistan Information Management Services (AIMS), United States Geological Survey (USGS), Ministry of Energy and Water (MEW), company Consultative Group on International Agricultural Research (CGIAR), World Wide Fund (WWF), Food and Agricultural Organization (FAO) etc. All the data is stored as geo-database was shown above the structure hierarchically.

V. Methodology

It is a perfect way of considering the information and data on Kabul basin water resources development and assessing potential dams sites in the basin and focusing on the following issues.

a) Literature reviewed on the current environment in the Kabul Basin on the main parameters influencing dam site location methodologies for dam site assessment with a focus on GIS techniques.

b) Thorough understanding of the existing methodology in order to define the processes involved and to assess the feasibility of such activity in the country (data availability, GIS techniques required).

These included:

- GIS methodology for initial site screening using terrain data followed by more detailed assessment of the sites obtained from the generic terrain screening
- GIS also used for water resources assessment of the future available water in the dam's catchment area
- Procedures that took into account thematic maps, base maps and other detailed maps has been generated for the basin and definition of characteristics of dam site location assessment.
- Asses availability of required geological data layers assist with assessing suitability of the underground for dam implantation.
- Geo-database and knowledge base management and development.

c) FUNCTIONALITY OF AR CHYDRO

The Arc Hydro Toolset is a suite of tools which facilitate the creation, manipulation, and display of Arc Hydro features and objects within the ArcMap environment. The tools provide raster, vector, and time series functionality, and many of them populate the attributes of Arc Hydro features [4].

1) Terrain preprocessing

Terrain Preprocessing uses DEM to identify the surface drainage pattern. Once preprocessed, the DEM and its derivatives can be used for efficient watershed delineation and stream network generation.

- Level DEM: assign constant elevations under lake polygons
- DEM Reconditioning: burn in existing streams
- Build Walls: burn in existing boundaries
- Fill Sinks: This function fills the sinks in a grid. If a cell is surrounded by higher

Elevation cells, the water is trapped in that cell and cannot flow. The Fill Sinks function modifies the elevation value to eliminate these problems.

2) Terrain Processing

Terrain Processing - the functions that will create the data supporting the delineation process:

- Adjust Flow Direction in Lakes
- Flow Accumulation - computes the flow accumulation grid that contains the accumulated number of cells upstream of a cell, for each cell in the input grid
- Stream Definition: This function computes a stream grid based on a flow accumulation grid and a user specified threshold.
- Stream Segmentation: This function creates a grid of stream segments that have a unique identification
- Catchment Grid Delineation: This function creates a grid in which each cell carries a value (grid code) indicating to which catchment the cell belongs.
- Catchment Polygon Processing
- Drainage Line Processing
- Adjoint Catchment: This function generates the aggregated upstream catchments from the "Catchment" feature class
- Drainage Point Processing: This function allows generating the drainage points associated to the catchments
- Longest Flow Path for Catchments

3) DEM Pre-processing

There are four key elements that define the expected "behavior" of the flow patterns in the terrain:

a) Sinks (depressions, pits). Sinks are the areas into which the water flows but does not exit as surface flow. In DEMs, most of the sinks are artificial and are artifacts of DEM construction. There are also real sinks. Sinks can be a function of the analysis. For low flow conditions, some sinks will capture water that will never leave the sink and will not contribute downstream, while under high flows, they will fill and spill over the sink boundary and eventually contribute to the flow downstream.

b) Known streams: Known streams represent observed drainage patterns captured as a vector polyline layer. The expectation is that the drainage pattern generated by the DEM will match the drainage pattern represented by the vector layer.

c) Known lakes. Known lakes represent observed lakes captured as a vector polygon layer. Lakes can be either sinks, where all the water drains into the lake and none comes out, or they can have an outlet stream (in which case the water entering the lake will exit through the stream draining the lake).

d) Known drainage area boundaries. Known drainage area boundaries represent known boundaries captured as vector polygon layers. Any "droplet" of water will stay within the drainage boundaries and drain either to the sink within the drainage area or to one drainage area outlet point [5]. DEM Pre-processing with hydrological is a part conducted in Arc hydro tool as the data supporting process:

- DEM Reconditioning- This function modifies a DEM by imposing linear features onto it (burning/fencing). The function needs as input a raw dem and a linear feature class (like the river network) that both have to be present in the map document.

- Fill sinks- This function fills the sinks in a grid. If a cell is surrounded by higher elevation cells, the water is trapped in that cell and can not flow. The Fill Sinks function modifies the elevation value to eliminate these problems, The output is the Hydro DEM layer, named by default Fil.
- Flow Direction- This function computes the flow direction for a given grid. The values in the cells of the flow direction grid indicate the direction of the steepest descent from that cell.
- Flow Accumulation- Computes the flow accumulation grid that contains the accumulated number of cells upstream of a cell, for each cell in the input grid.
- Stream Definition- This function computes a stream grid based on a flow accumulation grid and a user specified threshold. The cells in the input flow accumulation grid that have a value greater than the threshold are assigned a value of 1 in the stream grid. All other cells are assigned no data.
- Stream Segmentation- This function creates a grid of stream segments that have a unique identification. Either a segment may be a head segment, or it may be defined as a segment between two segment junctions. All the cells in a particular segment have the same grid code that is specific to that segment.
- Catchment Grid Delineation- This function creates a grid in which each cell carries a value (grid code) indicating to which catchment the cell belongs. The catchments correspond to your river segments – the more segments you have, the more catchments will be generated. The value corresponds to the value carried by the stream segment that drains that area, defined in the stream segment link.
- Catchment Polygon processing- Converts the raster data developed so far to vector format. The rasters created until now have all been stored in a folder named *Layers*. *The vector data will be stored in a feature dataset also named Layers within the geodatabase associated with the map document.*
- Drainage line processing- This function converts the input Stream Link grid into a Drainage Line feature class. Each line in the feature class carries the identifier of the catchment in which it resides.
- Drainage point processing- This optional function allows generating drainage points associated with individual catchment.

4) Watershed processing

Arc Hydro toolbar also provides an extensive set of tools for delineating watersheds and sub- watersheds. These tools rely on the datasets derived during terrain processing this has been done for delineating sub watersheds for existing points (eg. gaging sites) and batch watershed delineation.

VI. Simulation Experiment

A. kabul river basin regional setting (indus basin)

Briefly in terms of regional setting illustrates the extent and location of the KRB within the Indus Basin, that covers Afghanistan, India and China as well as a larger part of Pakistan. The basin has a total drainage area exceeding 1,165,000 km2 and it has been estimated annual flow stands at around 207 km3.

The Indus Basin comprises of the Indus River, its five major left bank tributaries the Jhelum, Ravi, Chenab, Sutlej, and Beas rivers and one of major right tributary the Kabul River Fig. 3.

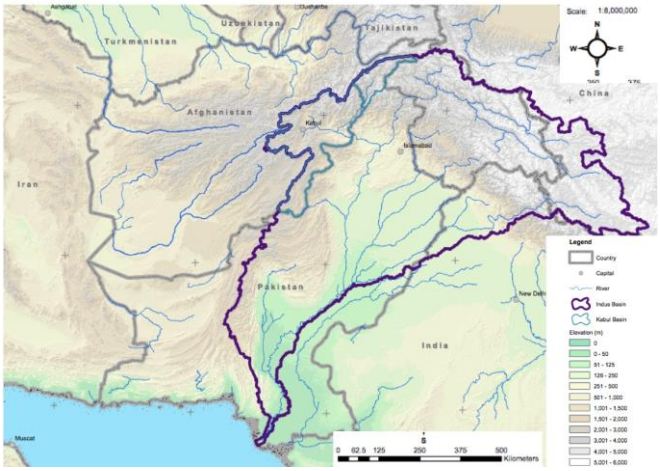


Fig. 3. The regional setting map on Kabul River Basin

B. Topography And Elevations

The elevation presented in this map is from SRTM version4 DEM dataset in combination with shaded relief overlap for presentation enhancement Fig. 4. The river system/ network has been computed form the same DEM through hydrological modeling.

The elevation map highlights the relief of the Kabul Basin and illustrates the shape of the land surface its topography. The elevation, in combination with land cover, surface roughness, and soil characteristics is the most necessary factor defining the hydrological characteristics of the river basin and its drainage system.

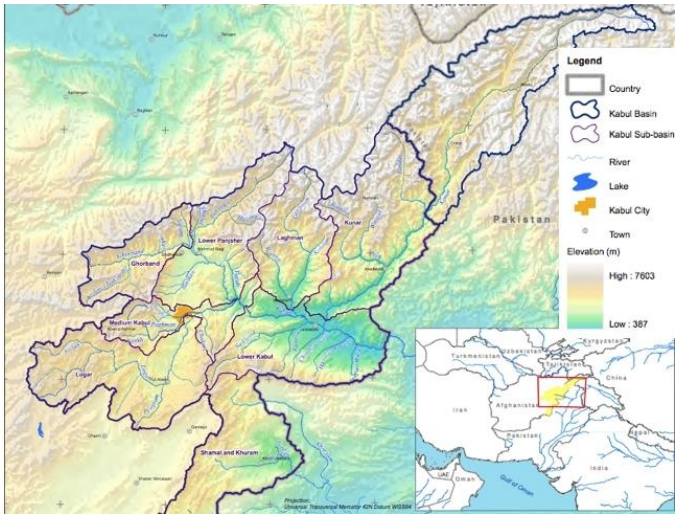


Fig. 4. Presents Kabul basin topography

C. River classification and codification

Kabul River left bank and right bank tributaries classified and codified in attribute the river coding creates a unique codification support for river and hydrological stations and allows linking the GIS layers with other data in the KB. A river code number consists of three groups:

- The first group relates to the main river basin in which the river lies.
- The second group stands for first order tributary.
- The third group generally consists of three digits and stands for various "second order" and "third order" tributaries listed in serial order, and a latter "L" or "R" for left bank tributary and right bank tributary respectively.

The Fig. 5, provides a schematic diagram of the Kabul river and its main tributaries – including naming and coding system.

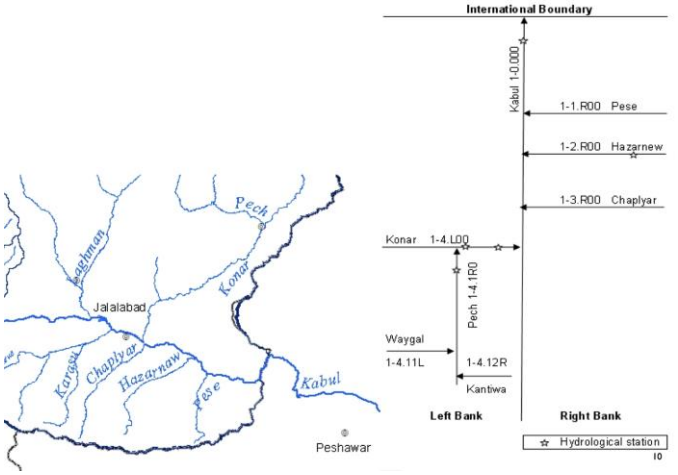


Fig. 5. Shows rivers classification and codification

D. Kabul administrative units

The river and sub-basins map present the river system and watersheds in the Kabul Basin. Basin and sub-basins are natural stream drainage areas used for collecting and organizing hydrologic data.

GIS layers have been consisted of their feature and attributes, this map presenting Kabul Basin administrative units, the administrative divisions of Afghanistan is districts and provinces respectively. There are 34 provinces in the country, and province is further divided into districts in total there are 127 districts in the Kabul basin Fig.6, This basin has been covering ten provinces such as Kabul, Wardak, Logar, Nangrahar, Kunar, Nuristan, Laghman, Kapisa, Panjsher, Parwan as well as some part of Ghezni province. Except for the Ghazni province and Wardak, the basin boundaries correspond to the provincial boundaries.

Other four basins of the country (Panj Amu, Northern river basin, Harirud Murghab river basin, Helmand river basin) cover the remaining provinces and districts.

Fig. 6. Illustrates Kabul basin administrative divisions

E. river infrastructure and irrigation

Using GIS application plays a significant role in river basin management and dams site selection as well as water infrastructure projects (existing, planed and being studied) in the basin based on raster and vector required data.

The irrigated areas have been estimated from satellite image interpretaion and classification using Landsat imageries presents the current situation of Kabul River Basin Afghanistan.

The most important water uses in the KRB included irrigation, energy generation, and industrial as well as uses for domestics. In Fig. 7, map illustrates the location of water infrastructure existing and planned projects and also dams site locations as well as the main urban centers and irrigated areas in the basin. In addition, representations of these layers can also be found in the more detailed sub-basins maps.

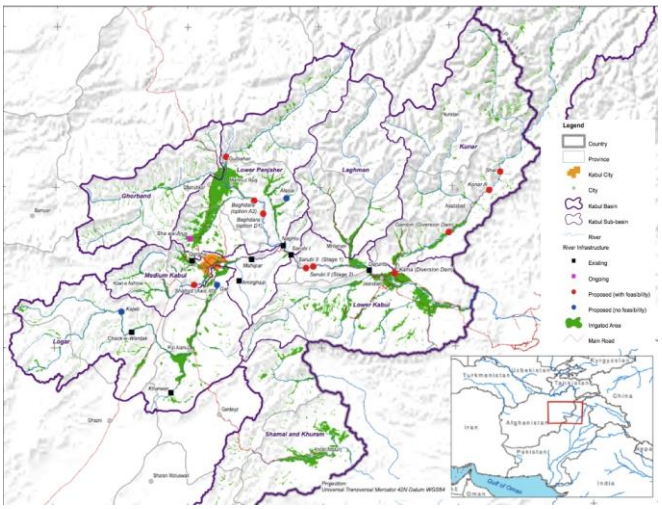


Fig. 7. Presents water infrastructure and irrigation of Kabul basin

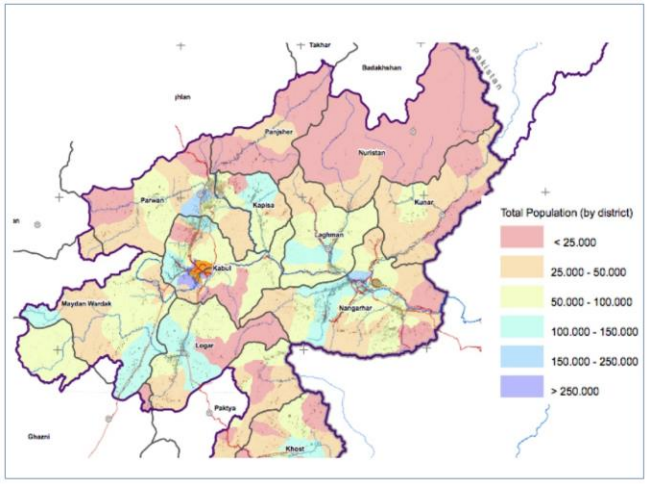


Fig. 8. Shows total Population density by district in Kabul basin

VII. THEMATIC MAPS

As mentioned before in terms of the name of some thematic maps that have been produced for realizing the basin characteristics and features. Thematic map emphasizes a particular theme or special topic such as the average distribution of rainfall in an area as we as meteorology or hydro-meteorological of an area.

They are different from general reference maps because they do not just show natural features like rivers, cities, political subdivisions and streets. Instead, if these items are on a thematic map, they are simply used as reference points to enhance one's understanding of the map's theme and purpose.

A. population density

This dataset has a 1 km resolution (30" X 30"), and is the finest resolution global population distribution data. The algorithm uses spatial data and imagery analysis technologies and a multi-variable asymmetric modeling approach to disaggregate census counts within an administrative boundary [7].

GIS has used for the analyzing population density based on the required GIS data for the basin, statistic data has been gotten from the Central Statistic Office (CSO), while the raster dataset shown is from the Landscan dataset.

The population map in Fig. 8, presents the total population by district in the Kabul. The latter dataset illustrates the areas with higher population density, and confirms the strong correlation of population density with irrigated areas along the river systems.

B. Hydro-meteorology of the basin

The hydrographs show the flow rate (discharge) versus time past a specific point in a river. The rate of flow is typically expressed in cubic meters per second.

River flow data is one of the main input data for effective river basin management. In surface water hydrology, a Hydrograph is a time record of the discharge of a stream, river or watershed outlet. Rainfall and/or snowmelt is typically the main driver of watershed discharge and a hydrograph records how a watershed responds to these and other drivers of watershed hydrology [8]. For instance Fig. 9, presents the graph of hydrological station of Tangi Saidan on Kabul River let to present this station point as it's been take into consideration and generated from the historical available data using GIS tool.

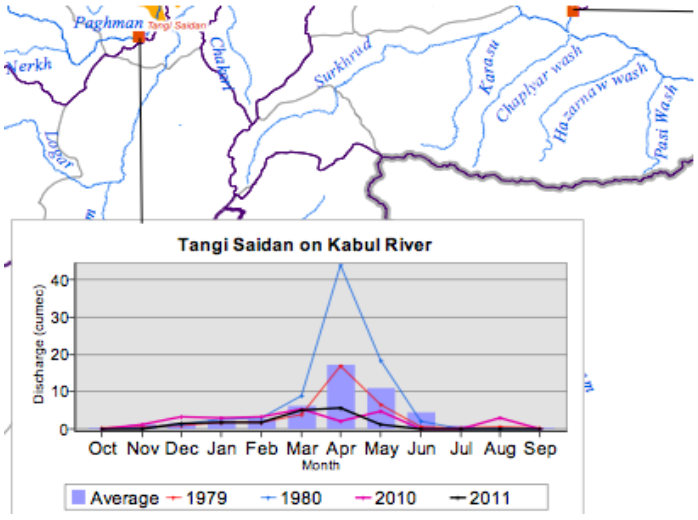


Fig. 9. Shown discharge of Tangi Saidan in cubic meter

C. Precipitation and temperature

This map in Fig. 10, presents the average yearly precipitation (rainfall and snow) over the Kabul Basin. Precipitation variability is extreme in the basin and in strongly correlated with the elevation. Precipitation ranges from 200mm or less around Jalalabad (semi-arid climate) up to 3000 mm or more in the high altitudes in the northern part of the basin (highland climate) [9].

The graphs near the show the monthly precipitation and temperatures in a few selected locations, where a climate station is located.

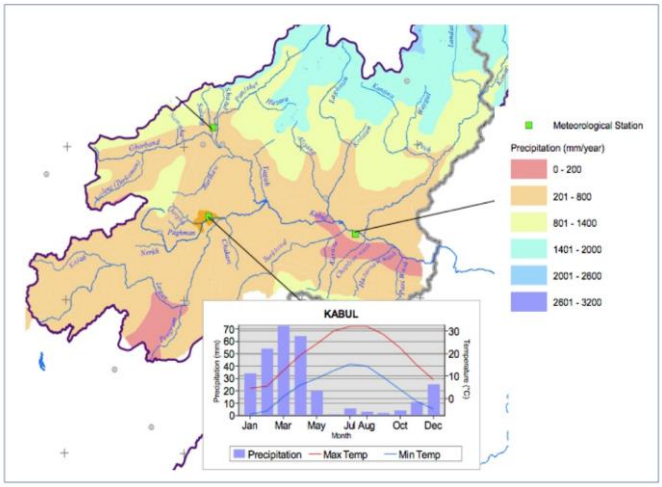


Fig. 10. Shown precipitation and temperature of Kabul basin

This data is very important for water resource infrastructure development and also for knowledge and knowing about the features of the basin for implementation of projects not only for water resources but also for other projects such as irrigation and water supplement projects as well as for investment plan in the basin critically needed either large and medium water infrastructure projects. The population has been increasing rapidly in the basin and water demand and economic growth demand increase.

ACKNOWLEDGMENT

I am grateful for Japan International Cooperation Agency (JICA) Project For the Promotion and Enhancement of the Afghan Capacity for Effective Development (PEACE) and University of the Ryukyus for the given opportunity and their financial support, so the product of this International Journal paper would not be possible without them.

REFERENCES

[1] Akbari, M.A, Tahir, M,Litke, D.W and Chornack, M.P, “Ground-water levels in the Kabul basin, Afghanistan”, U.S Geological Survey Open-file report, pp. 46, March 2007.

[2] Thomas J. Michael P. Chonack, Mohammad R. Taher “Ground Water-Level trends and implications for sustainable water use in the Kabul Basin, Afghanistan,” Afghanistan Geological Survey, pp.457-467, September 2013.

[3] Ali Ershadi, Hamid Khiabani, Jens Kristain Lorup “applications of Remote Sensing, GIS and River Basin Modelling in integrated Water Resource Management of Kabul River Basin”, TOOSS AB consulting Engineering, Mashhad, Iran. pp. 10, May 2005.

[4] Rasooli Ahmadullah, Kang Dongshik,"Assessment of potential dam site in the Kabul basin using Geographical Information System," University of The Ryukyus, Japan, International Technical Conference on Circuits/Systems, Computers and Communications (ITC-CSCC) Phuket, Thailand, pp. 521-524, July 2014.

[5] Sabbar Abdulla Salih, Abdul Salam Mehdi Al-Tarif, “Using of GIS Spatial Analyses to Study the Selected Location for dam reservoir on Wadi Al-Jirnaf, west of Shirqat area, Iraq” University of Tikrit, Iraq, pp. 117-127, January 2012.

[6] Mudher N. Abdulla, Senior Chief Engineer, MSc in photogrammetry & RS, Expert in GIS &RS, MOWR – IRAQ, “Catchment Area Delineation Using GIS technique for Bekhma Dam”, pp.18.May 2011.

[7] Anwar M.M.[1], Bhalli M.N.[2] "Gojra, Pakistan. urban population growth monitoring and land use classification by using GIS and remote sensing techniques acase study of faisalabad city", Department of Geography, GC University Faisalabad, & [2]Department of Geography, Govt. Postgraduate College, February 2012.

[8] Sustainable Development Department South Asia Region WB, "Scoping Strategic Options for Development of the Kabul River Basin". A Multisectoral Decision Support System Approach 2010 Afghanistan, 1818 H Street, NW, Washington, D.C. 20433 USA.

[9] Aquastat, Afghanistan, Food and Agriculture Organization (FAO) of the United Nation, "Geography, climate and population Geography", 2012.

The Examination of Using Business Intelligence Systems by Enterprises in Hungary

Peter Sasvari
Institute of Business Sciences, Faculty of Economics
University of Miskolc
Miskolc-Egyetemvaros, Hungary

***Abstract*—Data are one of the key elements in corporate decision-making, without them, the decision-making process cannot be imagined. As a consequence, different analytical tools are needed that allow the efficient use of data, information and knowledge. These analytical tools are commonly called Business Intelligence systems that are introduced into the opeartion of enterprises to make access to business data easier, faster and broader in line with the needs of a given enterprise. Based on the findings of an empirical survey, this paper aims to give a deeper insight of the causes and purposes of using BI systems by Hungarian enterprises. It is revealed that such systems are mostly used for risk analysis, financial analysis, market analysis and controlling while their potential to make predictions is usually overlooked. One important conclusion of the paper is that the faster spread of BI systems would be facilitated by reducing costs, simpler parameter settings and a higher level of data protection.**

Keywords—Business Intelligence; Hungary; Enterprises

I. INTRODUCTION

Nowadays a revolution is taking place in the field of information technology. The globalized and fast-paced business world together with a lot of other factors brings about a continuous increase in the volume of data [15]. Businesses have to discover the knowledge in the depths of overflowing data if they want to be successful. A fast and flexible response is an indispensable condition for a company's advancement. Enterprises should use huge amounts of data in a way that helps them to make profits sooner or later [5]. The significant amount of unstructured data is not sufficient to allow the company's management to make the most advantageous decisions in certain situations. As a first step, data must be converted into information and then into profitable knowledge, Knowledge through which the management can make the most profitable strategic decision for the organization. Thus, application of the business is essential to stay ahead of tools and technologies that are affected by these exploratory processes.

First, a brief review is given on the definition and concept of Business Intelligence systems, then the main elements and the advantages of using BI systems are presented. Based on an empirical research carried out in Hungary, the benefits of using such systems are demonstrated through the experience of Hungarian enterprises. Finally, some thought are given to the question of what factors could contribute to the faster penetration of BI systems.

II. THE FIRST CONCEPTS OF BI

In vain has become one of today's commonly used in an expression of BI, not a definition can be easy task. And a number of different conceptual approaches, a description can be found in the exploration of the topic. The literature on the concept of the creation of Howard Dresner name connects who in 1989 had defined BI systems as "*in the broad category of software and solutions for gathering, consolidating, analysing and providing access to data in a way that lets enterprise users make better business decisions*".[17]

Cser, Fajszi and Feher [13] further work detailed approach can discover the BI determination: "The business intelligence technologies, applications, methodologies, total process management and organizational units that are prepared and business decision-making process of the company support throughout the whole corporate data assets." This formulation breaks the components of BI, while highlighting the most important function of the decision aid. A term used by Gabor et al [1] says: "Business Intelligence solutions build on the achievements of the knowledge management, data warehousing, data mining and business analytics. With the help of key processes, identify data tracking, it becomes possible to improve; identify, monitor trends within the company, competitors and market performance." This document mainly focuses on using BI as practicable activities and processes. Loshin [4] offers another definition which states: "The necessary processes, technologies and tools for conversion into information, information into knowledge, and knowledge into plans forming that the drivers of profitable business. Business Intelligence includes data warehousing, business analysis tools, and knowledge management." This concept sees BI as a process by which this data is actually information that can be further converted into knowledge to ensure profitability. Sabherwal [21] defines BI as providing decision-makers with valuable information and knowledge by leveraging a variety of sources of data as well as structured and unstructured information.

III. THE COMPONENTS OF BI

Main elements of the BI are pronounced in the word investigation since the existence and co-operation of these components is essential to companies that actually take advantage of the benefits. The components are the following [23]: data mining; source of data; data warehousing; data visualization; decision support; Online Analytical Processing (OLAP).

Data source is called source data warehousing that a variety of sources, the system may come from within the company [6]. We distinguish between internal data, which are formed within the organization and external data. The external data source may be economic environment, competitors or data on clients [6]. The BI concept is often used in conjunction with the data warehouse [2] concept, which is subject-oriented, integrated, data historicity storage, durable system whose main objective is to provide efficient information extraction of the data, in particular to support the decision-making process. The BI is an important component as well as the data visualization [16], which of those technologies is the common name, enabling the presentation of data and in some cases additional information was obtained in the data even according to their interpretation also some sub-processes in the data processing.

Effective knowledge mining, in order to gain advantages in a large amount of business -data found within an organization or from other sources of data mining [22]. OLAP [7], a software technology that enables analysts, businessmen and managers to company information organized according to dimensions assess compliance levels for rapid, consistent and interactive way. The one profession BI decision support [2], which is a combination of model-based data processing and decision-making processes that help managers decision-making activities.

IV. Advantages and Problems Arising from the Introduction of BI Systems

Businesses decide to use a system because it benefits from the operation of a business acquisition and positive impact on the company 's progress on waiting for it. Otherwise it is not for the BI. The BI application offers many benefits to the employing organization, since the data is forging business advantage. Such advantages may include [20]:

- **Data consistency check**: BI applications can test the correctness, consistency of the data. The consistency of a database refers to the relationship between the number of data, where the content of a data element occurrences of the same as that of another of the same occurrence.
- **Reduction of data redundancy**: By redundancy we mean a data store more than one place in a database. It is difficult to avoid redundancy does not occur, but multiple occurrences to minimize any case an important goal. This can be an important tool for BI systems [24].
- **Fast and robust decision-making**: Through the BI decision makers of companies from a variety of business areas to look through the detailed and updated analysis due to which their decisions can help immediately and established facts may be taken.
- **Access Faster Information**: BI applications allow the necessary information there and then they have access to the users when and where you are need it. This will ensure fast and efficient information management.
- **Making Effective Predictions**: Based on the companies collected historical data, future environmental characteristics make an effective and reliable forecasts based on some estimated values.
- **Improved internal communication**: Enables organizations from within the enterprise more efficient communication, as it not only makes available the key users of the data.
- **Improved data security**: By achieving organizational data from a single portal is made, which is protected by the business - intelligence data security passwords properly treated.
- **Exceptions Exploration**: Includes corporate fraud, claims to detect. The significant advantage of some BI applications to existing applications and in business rules effectively combines the intelligent statistical forecasting models, and these together with the use of fraud detection.
- **Reduced costs**: for example, the drop in labor costs, costs of the time needed (eg, lower IT operating costs) and the costs of manual processing of any mistake by restoring the BI application.
- **Involvement of mobile devices**: gate opens towards you using BI to mobile devices with the involvement of staff from the office is able to carry out their duties.
- **Handling of larger amounts of data**: A massive amount of enterprise BI solutions for data collection, analysis and management can do.
- **Increased profits**: Businesses can gain more profits due to the BI technology , as these applications provide for them to receive relevant information about the service and reap the benefits derived from their customers and the market as a profitable investment before acquiring their competitors [14].
- **More accurate stock records**: BI software inventory monitoring of assisting companies in providing the right amount of stock available to push when the customer needs it and not let the company build up inventory.
- **Revealing Hidden Business Rules**: The BI system business strategies, the company is known and viewing of hidden rules for development and change management [9].
- **To achieve more relevant information to BI applications more access to information about new learning opportunities, idea, services and products.** It will also help the company to choose its own leaders. In addition, the reliable and timely information to help businesses make the right information to make informed decisions. This information may include any topic, such as agriculture, fisheries, land, education, vocational skills, livelihood, economy, etc.
- **Customer needs, behaviour understanding and reaching a wider customer base: BI applications collect information on customers and clients.** So businesses can learn about their customers' needs, adapt to their needs. Learn about current and future customer’s habits and build up a wider range of customers.

- **Exploring patterns of behaviour**: The behavioral economics of the economic behavior of individuals engaged in the researching drivers of individual decisions and actions. Revealed by behavioral economics are different rules often assumed by neoclassical economics, behavioral rules. It is not yet clear to what extent compatible with these different sets of rules, such that it is not completely clear how revealed by behavioral economics behavioral patterns affect social outcomes, ie. what effect on the aggregate level [8].

BI projects often face obstacles during construction. The reasons are many. Based on Santané et al [6], the following factors can be traced to the failure of any BI project. (1)The project does not receive the necessary business support. (2) The project members are not sufficiently skilled and/or are not available; sometimes its participants can not be used for certain tasks properly. (3) The project participants are not willing to participate actively in the tasks. (4) Not enough attention is paid to issues related to data quality (not taken into account the effect of the poor quality of data in business profitability). (5) Not used in software development methodology. (6) Improper handling, use of metadata. Metadata is "*data about data*" [19]. (7) Is the organization aware that a BI project, the department spanning initiative so different from the independent solutions used previously. (8) You can not perform business analysis. (9) No work breakdown structure. (10) Too insisting on a specific methodology or tool.

V. THE RESEARCH METHOD

Examination of the characteristics of the research is to use BI applications operating in the Hungarian enterprises. The research has been carried out by a questionnaire. The questionnaires were sent out to businesses randomly, regardless of company size, activity and geographical location. The filling questionnaire of a sample is not representative, so the results of the investigation only interpreted the scope of companies involved in the research.

A. The objectives of the research

The primary objective of the research is examining the experiences of BI applications among Hungarian enterprises. The research aims to

1) Uncover the basis of the responses received from the first questionnaire to a different company sizes are expected to benefit from the business use of BI systems ,

2) Companies to perform tasks that used for BI applications

3) Success or failure of the third company size varies by whether the use of BI.

Among the objectives included assessment of the future use of BI technologies, taking into account the categories of enterprises, and in the investigation as to how the business of the opinion that it should develop a greater proliferation of BI systems.

B. The research assumptions

The first step of the research assumed examines the benefits of using the BI enterprise size. In this connection, the following assumptions were made.

1) Category in different advantages in the use of BI.

2) It is more common for corporations and medium-sized enterprises in the positive experience than small businesses in connection with the operation of BI applications. (a) Perform different size categories of tasks used in BI applications. (b) Areas of successful BI applications using the different enterprise size. (c) Using the areas of BI applications unsuccessful vary by enterprise size.

3) Size categories are different opinions about how it should develop BI applications.

C. The model features

The questionnaire was filled out by 77 pieces of Hungarian enterprises in 2013. Examining the binding of the companies that completed size concluded that the sample represent themselves in a similar proportion of micro, small and medium-sized enterprises and large corporations.

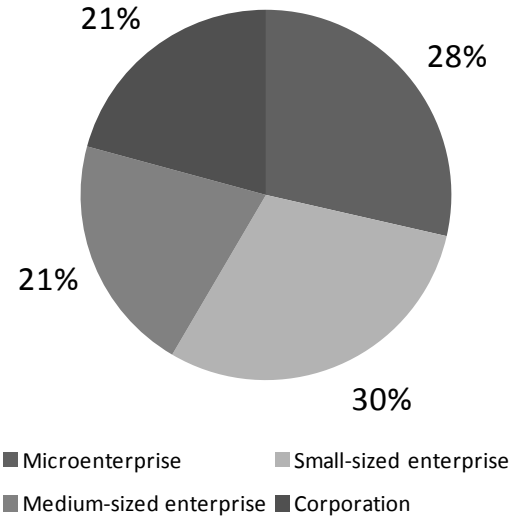


Fig. 1. The distribution of respondents by size categories

Less than one tenth of the respondents said that they work in the enterprise BI system or application. The studied companies are large companies have most BI applications, small and medium-sized enterprises is almost the same percentage, while the micro-enterprises are not used at all. Determining the proportion of those firms that do not operate the system and no plans to introduce them.

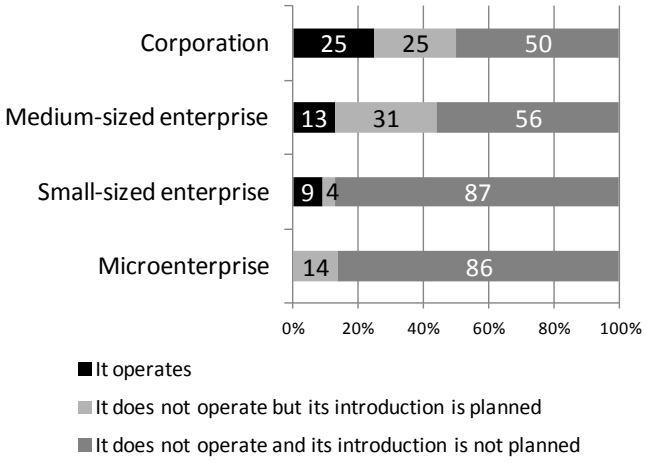


Fig. 2. The application of BI systems in Hungary by company size categories in 2013

VI. The Frequency of BI Applications Among the Hungarian Enterprises

According to IDC [11] in 2011 to assess the Hungarian BI applications market growth characteristics. An interesting result of the study, despite occurring in 2008, the global economic crisis in 2009 increased by 14.7% in the Hungarian BI market size. In the coming years, the market also increased, but to a lesser extent. The 2011 analysis predicts that in 2013, an increase of 7% will be [10]. A significant part of the query and analytics market makes up the other part is made up of advanced analytical applications [24].

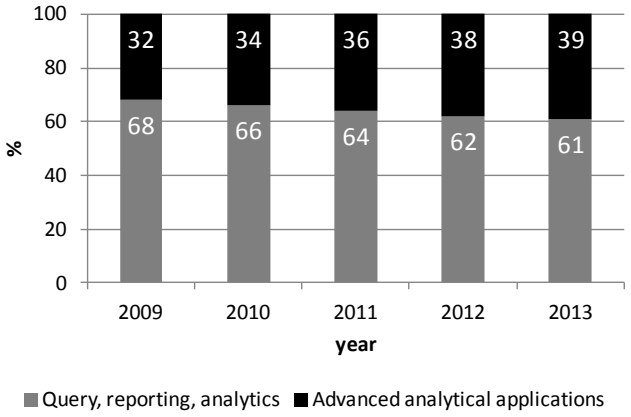


Fig. 3. Changes in the size of the Hungarian Business Intelligence market

Based on the survey of the Hungarian small and medium-sized enterprises hardly, while micro enterprises did not apply the BI applications in their activities. The large companies surveyed only one fifth to one tenth of the small and medium-sized businesses have been in possession of a BI application, but all three major categories of businesses, the percentage of people who do not plan to use these systems in the future [23].

A. Expected advantages of using BI systems

The companies aim to provide your company service systems are set to benefit from the use. Advantage, which may be more favorable position in the market over their competitors. The company size categories, each handling a larger amount of data clearly waiting for BI applications [19]. The use of hypothetical positive effect also seen as a quicker access to information, improved and more accurate inventory records of internal communications.

TABLE I. Expected Advantages of Using BI Systems in Hungary by Size Categories in Hungary in 2013

Activity	Micro-enterprise /ranking/	Small-sized enterprise /ranking/	Medium-sized enterprise /ranking/	Corpo-ration /ranking/
Handling greater amount of data	1 (64%)	1 (83%)	1 (75%)	1 (69%)
Faster access to information	2 (59%)	2 (57%)	3 (44%)	2 (56%)
More appropriate inventory	3 (50%)	5 (48%)	2 (56%)	6 (25%)
Forecasting	4 (41%)	3 (52%)	7 (31%)	3 (50%)
Improved internal communication	6 (41%)	4 (52%)	4 (44%)	4 (44%)
Access to a broader range of customers	5 (41%)	7 (22%)	5 (38%)	7 (25%)
Access to more relevant information	7 (36%)	6 (26%)	8 (31%)	5 (31%)
Data consistency control	9 (18%)	9 (17%)	6 (38%)	8 (25%)
Better ability to adaptation	10 (14%)	10 (13%)	9 (25%)	11 (13%)
Higher profits	13 (9%)	8 (22%)	12 (13%)	10 (25%)
Reducing redundant data	8 (32%)	13 (9%)	11 (19%)	12 (13%)
Inclusion of mobile devices	14 (9%)	12 (13%)	10 (25%)	9 (25%)
Lower IT operation costs	12 (9%)	11 (13%)	13 (6%)	13 (13%)
Revealing exceptions	11 (9%)	14 (9%)	14 (6%)	14 (13%)
Revealing hidden business rules	15 (5%)	16 (0%)	15 (6%)	15 (13%)
Revealing behavioural patterns	16 (5%)	15 (9%)	16 (0%)	16 (6%)

Separately examining the size in the responses it can be seen that each of the four categories of handling large amounts of data response options marked the biggest percentage. In the case of micro and small enterprises and large corporations in second place for faster access to information is to be considered the benefits of BI, while medium-sized enterprises by analyzing the responses of more accurate inventory records occupies this space. Microenterprises in terms of more accurate set of records in third place. Improvement of internal communication for small and medium-sized enterprises classified in this position, the order of the projections of large and small businesses. The medium-sized enterprises percent of responses came from the same access to information faster response option such as communications improve. On completion of the chi- square test the result available to no one answers are significantly related to company size. Businesses of all sizes benefit from the same category BI application of the first three places are implied. So we can conclude that

company size has no effect on what the company expects to benefit from business intelligence applications.

B. Experiences gained on using BI systems in Hungary

BI applications are used to perform certain tasks in the organization, in order to speed up, facilitate and simplify work processes. Using BI in the enterprise may experience both positive and negative. This part of the research of micro-enterprises are not included, as they have stated they do not use BI systems so they could get no experience.

TABLE II. EXPERIENCES GAINED ON USING BI SYSTEMS IN HUNGARY BY SIZE CATEGORIES IN 2013

	Small-sized enterprises	Medium-sized enterprises	Corporations
All successful	6%	0%	0%
Rather successful	56%	75%	85%
Rather unsuccessful	0%	0%	0%
All unsuccessful	0%	0%	0%
No such projects	38%	25%	15%

Small businesses are 38% and 25% of medium-sized and corporations 15% of said that there was no organization either of these projects. That all experiences have been successful only on small companies arrived by a marginal response. Among the respondents "rather successful" response option was the most common, as all the three categories identified a significant proportion of this option. 56% of small businesses, medium-sized companies and three -quarters the size of the largest 85% of the company deems more successful BI functionality. The reason may be that the larger the business the more favorable economic situation, which allows input of larger amounts of the BI field. Qualified professionals and consultants can apply the right skills who are able to derive a BI project, thereby increasing the chances of successful implementation and use.

1) Activities performed by BI systems

Based on the responses received, it can be said that small and medium-sized businesses in the first place to carry out financial analysis applied to the BI tools in large companies, however, a striking degree of strategic decision-making. The company size secondly, market analysis, preparing and controlling the field are turning to BI support. The results of the study are summarized in the table below.

TABLE III. ACTIVITIES PERFORMED BY BI SYSTEMS IN HUNGARY BY SIZE CATEGORIES IN 2013

Activity	Small-sized enterprises		Medium-sized enterprises		Corporations	
	score	ran-king	score	ran-king	score	ran-king
Financial analysis	3.56	1	4.42	1	3 92	1
Controlling	3.44	2	4.08	2	3.92	2
Market analysis	3 .18	3	3 .83	3	3.23	5
Strategic decision-making	2.71	10	3.67	4	3.69	3
Forecasting	3.06	5	2.83	10	3.38	4
Optimization of transaction processes	3.00	6	3.33	7	2.92	6
Risk analysis	2.94	7	3.50	6	2.77	7
Product development	3.12	4	3.50	5	2.54	9
Customer relationship management	2.76	9	2.92	9	2.62	8
Activity monitoring	2.88	8	3.33	8	2.31	10

Performing the chi-square test showed no significant correlation with either case of company sizes.

2) Areas of the successful use of BI systems

Answering the survey questions for small businesses, according to the BI applications were in the field of financial analysis and market analysis are the most successful, and controlling, but also in relation to customer management success in the BI system. The medium-sized enterprises in particular have experienced success in controlling and monitoring activity. The second place in their case, the financial analysis, preparation of forecasts, market analysis and customer management replaced, and the third strategic decision-making. Set up by the big companies in the following order: financial analysis, controlling, strategic decision-making. The answers to the business as a whole experienced a case of risk analysis and product development in the least that the BI system would be successful.

TABLE IV. EXPERIENCES ON THE SUCCESSFUL IMPLEMENTATION OF BI SYSTEMS IN HUNGARY BY SIZE CATEGORIES IN HUNGARY IN 2013

Activity	Small-sized enterprises		Medium-sized enterprises		Corporations	
	fre-quency	ran-king	fre-quency	ran-king	fre-quency	ran-king
Financial analysis	26%	1	25%	3	56%	1
Controlling	17%	3	31%	1	50%	2
Market analysis	26%	2	25%	4	25%	4
Activity monitoring	9%	6	31%	2	6%	9
Customer relationship management	13%	5	25%	6	19%	6
Forecasting	17%	4	25%	5	25%	5
Optimization of transaction processes	9%	7	13%	8	13%	7
Strategic decision-making	4%	9	19%	7	31%	3
Risk analysis	0%	10	0%	10	13%	8
Product development	9%	8	6%	9	6%	10

The related false assumption, since cross-tabulations by running test can not be said that the company size affect the performance of certain tasks BI systems success.

3) Areas of unsuccessful use of BI systems

Examination of the experience gained in the use of BI systems in the study continued to fail in what were most of these applications, and this is in the context of company size if they fail to reach an area of BI.

Among the small businesses in the strategic decision-making was the least successful of BI systems. The medium-sized enterprises stated that they experienced the most failures in the field of risk analysis, and forecasts for the major companies. Based on the responses of the enterprises controlling the performance of the tasks was the least typical BI fails.

TABLE V. EXPERIENCES ON THE UNSUCCESSFUL IMPLEMENTATION OF BI SYSTEMS IN HUNGARY BY SIZE CATEGORIES IN HUNGARY IN 2013

Activity	Small-sized enterprises		Medium-sized enterprises		Corporations	
	frequency	ranking	frequency	ranking	frequency	ranking
Forecasting	9%	6	13%	5	31%	1
Optimization of transaction processes	9%	5	19%	2	6%	6
Product development	4%	7	6%	8	19%	3
Risk analysis	13%	2	25%	1	0%	9
Unsuccessful introduction of applications	13%	4	13%	4	13%	5
Market analysis	0%	11	19%	3	19%	2
Customer relationship management	4%	9	6%	10	19%	4
Strategic decision-making	22%	1	0%	11	0%	11
Unsuccessful development project	13%	3	6%	7	6%	8
Activity monitoring	4%	8	6%	9	0%	10
Financial analysis	0%	12	13%	6	6%	7
Controlling	4%	10	0%	12	0%	12

The assumption is that company size affects the BI fail to fulfill the various tasks. Examining the adequacy of the established claim in case of failure occurring in the strategic decision-making in a significant relationship between the size of the business. The smaller the business is, the more experience there is the BI system failure in carrying out this task.

C. The future use of BI systems in Hungary

The ever-accelerating world of BI tools, the use of technology becomes a necessity for businesses. Half of small business and large corporations believe that the spread of BI systems to help reduce costs the most. Small businesses are 65%, while 56% of medium-sized businesses have the same view. The Hungarian companies are reluctant to spend higher amounts on BI systems, which introduce the size of the enterprise, the IT sector and related developmental depending on several million forints can range from 20 to 30 million forints (Kövesdi 2011). Secondly, the survey respondents criticized the level of protection of BI applications, as the respondent medium-sized businesses and large enterprises, 38% of them considered that the data protection should be repaired to ensure a higher level. Micro and small-sized enterprises with more than 20% think the same way. The existence of a simple parameter setting is considered to be the predominant BI roll-out of business. A medium-sized businesses for over 40% believe that the parameterisation simplify support the advancement of BI applications, micro-businesses and large enterprises and 20%, while 26% of small business takes the same view. Corporations according to the manufacturer's independence and the greater spread of standard features also contribute to the spread of these systems.

TABLE VI. FACTORS HELPING THE SPREAD OF BI SYSTEMS IN HUNGARY BY SIZE CATEGORIES IN 2013

Activity	Micro-enterprises /ranking/	Small-sized enterprises /ranking/	Medium-sized enterprises /ranking/	Corporations /ranking/
Cost reduction	1 (50%)	1 (65%)	1 (56%)	1 (50%)
Higher level of data protection	2 (27%)	3 (22%)	3 (38%)	2 (38%)
Simpler parameterizing	3 (18%)	2 (26%)	2 (44%)	5 (19%)
Higher independence from manufacturers	4 (14%)	5 (9%)	4 (25%)	3 (25%)
Extended standard functions	5 (14%)	4 (13%)	6 (13%)	4 (25%)
Better automatization of data integration	6 (0%)	6 (9%)	5 (19%)	6 (13%)

The assumption - that are the same size category as the opinions about how it should develop business intelligence applications - has become reasonably the results obtained, the opinion of the show business conform to the test aspects.

VII. CONCLUSION

The research aims to present and analyze the application of BI Hungarian experiences among enterprise. The paper intends to find answers to the following questions:

1) Companies to manage most of the larger amount of data for faster access to information and the provision of accurate stock records see the benefits of BI. The expected benefits of independent evolution of the size of the company.

2) Major Hungarian companies and medium-sized enterprises frequent positive experiences, such as small enterprises in the use of BI, as a result of the analysis indicate that the larger the firm, the more positive experiences have enjoyed the use of the BI. (a) Businesses most of the risk analysis, financial analysis, market analysis and controlling field is applied to BI applications. They do this regardless of méretkategóriától. (b) The most important areas of successful BI use of financial analysis, market analysis and controlling independently from size categories. (c) To fail in the most important areas of BI Making use of the predictions, the introduction and application of risk analysis. Significant relationship between company size can be observed in the area of strategic decision-making.

3) Businesses to reduce costs, higher levels of data protection and simple parameterization is considered so that could be the key to the spread of BI systems.

REFERENCES

[1] A. Gábor, "Üzleti informatika", AULA Kiadó, Budapest, 2007

[2] A. Jánosa, "Üzleti intelligencia alkalmazások", ComputerBooks Kiadó, Budapest, 2010.

[3] B. Marr, "What is Business Intelligence (BI)?", Advanced Performance Institute In: http://www.ap-institute.com/Business%20Intelligence.html

[4] D. Loshin, "Business Intelligence: The Savvy Manager's Guide", Newnes, 2012.

[5] D. R. Szabó "Kockázatelemző szoftverek összehasonlító elemzése", in Kovács Norbert, Építőkockák. 199 p. Győr: Universitas-Győr Nonprofit Kft., pp. 113-136. 2014.

[6] E. Sántáné- Tóth, M. Bíró, A. Gábor, A. Kő and L. Lovrics, "Döntéstámogató rendszerek", Panem Könvkiadó, Budapest, 2008

[7] E.F. Codd, S.B Codd, and C.T. Salley, "Providing OLAP (On-line Analytical Processing) to User-Analysts: An IT Mandate, Codd & Date", 1993.

[8] G. Koltay and J.Vincze J. "Fogyasztói döntések a viselkedési közgazdaságtan szemszögéből", Közgazdasági Szemle, LVI. évf., 2009. június, 495-525. o.

[9] I. Graham "Business Rules Management and Service Oriented Architecture: A Pattern Language", Wiley, 2007, ISBN 978-0470027219

[10] "IKT helyzetelemzés 2010 végén", 2011 - See more at: http://einclusion.hu/2011-01-27/ikt-helyzetelemzes-2010-vegen/

[11] "International Data Corporation (IDC)" - See more at: http://idchungary.hu/hun/about-idc/company-overview

[12] J. Szabó "Adatokból üzleti előny? Az üzleti intelligencia alkalmazások tapasztalatai a magyarországi vállalkozások körében", Miskolci Egyetem, TDK-dolgozat, 2013.

[13] L. Cser, B. Fajszi, T. Fehér, "Üzleti haszon az adatok mélyén, Az adatbányászat mindennapjai", Alinea Kiadó, 2010.

[14] L. James, "Top 14 Benefits of Business Intelligence", - See more at: http://smartdatacollective.com/yellowfin/42423/yellowfin-top-14-benefits-business-intelligence-part-one

[15] M. Aranyossy, A. Nemeslaki and A. Fekó, "Empirical Analyis of Public ICT Development Project Objectives in Hungary" International Journal of Advanced Computer Science and Applications (IJACSA), 5(12), 2014. http://dx.doi.org/10.14569/IJACSA.2014.051206 - See more at: http://thesai.org/Publications/ViewPaper?Volume=5&Issue=12&Code=IJACSA&SerialNo=6#sthash.CQYfHNXW.dpuf

[16] M. Friendly, "Milestones in the history of thematic cartography, statistical graphics, and data visualization", 2008 - See more at: http://www.math.yorku.ca/SCS/Gallery/milestone/milestone.pdf

[17] M. Gibson, D. Arnott, I. Jagielska and A. Melbourne, "Evaluating the Intangible Benefits of Business Intelligence", Review & Research Agenda, Proceedings of the 2004 IFIP International Conference on Decision Support Systems (DSS2004): Decision Support in an Uncertain and Complex World, pp. 295-305. 2004

[18] NISO, "Understanding Metadata", NISO Press. ISBN 1-880124-62-9. Retrieved 5 January 2010.

[19] P. Sasvári, "Az üzleti információs rendszerek használatának jellemzői a magyarországi és ausztriai kis- és középvállalkozások körében", 2013, - See more at:: http://www.irisro.org/gazdasagtan2013januar/G436SasvariPeter.pdf

[20] R. Bellu, "Microsoft Dynamics GP", Wiley, John & Sons, Incorporated, 2008.

[21] R. Sabherwal, "Succeding with Business Intelligence: Some Insights and Recommendations" Cutter Benchmark Review, 7 (9): 5-15, 2007.

[22] U. Fayyad, G. Piatetsky-Shapiro, P. Smyth, "From Data Mining to Knowledge Discovery in Databases", American Association for Artificial Intelligence, 1996.

[23] W. W. Eckerson, "Extending the Value of Your Data Warehousing Investment", The Data Warehousing Institute - See more at: http://tdwi.org/Articles/2007/05/10/Predictive-Analytics.aspx?Page=2, 2010.

[24] Z. Kövesdi "Üzleti intelligencia megoldások a vállalkozásoknál", Infotér elemzés, 2011.

Sentiment Analysis Based on Expanded Aspect and Polarity-Ambiguous Word Lexicon

Yanfang Cao, Pu Zhang, Anping Xiong
Department of Computer Science and Technology
Chongqing University of Posts and Telecommunication
Chongqing, China

***Abstract*—This paper focuses on the task of disambiguating polarity-ambiguous words and the task is reduced to sentiment classification of aspects, which we refer to sentiment expectation instead of semantic orientation widely used in previous researches. Polarity-ambiguous words refer to words like" large, small, high, low ", which pose a challenging task on sentiment analysis. In order to disambiguate polarity-ambiguous words, this paper constructs the aspect and polarity-ambiguous lexicon using a mutual bootstrapping algorithm. So the sentiment of polarity-ambiguous words in context can be decided collaboratively by the sentiment expectation of the aspects and polarity-ambiguous words' prior polarity.At sentence level, experiments show that our method is effective in sentiment analysis.**

Keywords—polarity-ambiguous word; aspect; sentiment analysis

I. INTRODUCTION

In recent years, sentiment analysis became a hot research topic in the field of natural language processing due to its widely use. Previous work on sentiment analysis has covered a wide range of tasks, including polarity classification, opinion extraction[1], and opinion source assignment.

One fundamental task at word level for sentiment analysis is to determine the sentiment orientations of words. There are basically two types of approaches for word polarity recognition: corpus-based and lexicon-based approaches. Corpus-based approaches using constraints on the co-occurrence of words and statistical measures of word association in the large corpus to determine the word sentiments[2]. On the other hand, lexicon-based approaches use information about lexical relationships and glosses such as synonyms and antonyms in WordNet to determine word sentiments based on a set of seed polarity words.

Overall those two methods aim to generate a large static polarity word lexicon marked with prior polarities out of context. In fact, a word may indicate different polarities depending on what aspect it is modified, especially for the polarity-ambiguous words, such as "高|high"，which has a positive orientation in snippet "high quality" but a negative orientation in snippet "high price". Though the quantity of polarity-ambiguous words is not large but polarity-ambiguous words cannot be avoided in a real-world application[1]. Unfortunately, polarity-ambiguous words are discarded by most research concerning sentiment analysis.

In this paper, the task of disambiguating polarity-ambiguous words is reduced to sentiment expectation of aspects. The sentiment expectation of aspects divide into two categories: positive expectation and negative expectation. A mutual bootstrapping algorithm is proposed in this paper to automatically construct the aspect and polarity-ambiguous words, utilizing relationships among aspects, polarity words and syntactic analysis. This algorithm is firstly initialized with a very small set of polarity-ambiguous words and syntactic patterns to retrieve a set of aspects. Then the sentiment expectation of an aspect is inferred utilizing the relations between aspects and polarity-ambiguous words in annotated reviews. Secondly, more polarity-ambiguous words is retrieved, utilizing the relations between aspects, syntactic patterns and annotated reviews. Finally, more syntactic patterns which are syntactic relations between aspects and polarity-ambiguous words, is retrieved. After several iterations the aspect and polarity-ambiguous word lexicon is constructed. Then the sentiment of polarity-ambiguous words in context can be decided collaboratively by the sentiment expectation of the aspects and the prior polarity of polarity-ambiguous words. At sentence level, experiments show that our method is effective.

II. RELATED WORK

Recently there has been extensive research in sentiment analysis and a large body of work on automatic SO prediction of words [2], but unfortunately they did not consider the SE of nouns in their research and regarded most of the nouns as "neutral". Some studies try to disambiguate the polarity of the polarity-ambiguous word [3]. Some researchers exploited the features of the sentences containing polarity-ambiguous to help disambiguate the polarity of the polarity-ambiguous word. For example, intra-sentence conjunction rule in sentences from large domain corpora is taken into consideration. Many contextual information of the word within the sentence is taken into consideration, such as exclamation words, emoticons and so on [4]. In order to automatically determine the semantic orientation of polarity-ambiguous word within context, some researches reduce this task to sentiment classification of target nouns, by mine the Web using lexico-syntactic patterns to infer sentiment expectation of nouns, and then exploit character-sentiment model to reduce noises caused by the Web data [5]. A bootstrapping method to automatically discover CPs and predict sentiment expectation of nouns is proposed by Wu in order to improve both sentence and document level sentiment analysis results [6].

The disambiguation of polarity-ambiguous words can also be considered as a problem of phrase-level sentiment analysis. For example, analyze its surrounding sentences' polarities to disambiguate polarity-ambiguous word's polarity in the sentence[7]. Use a holistic lexicon-based approach to solving the problem by exploiting external evidences and linguistic conventions of natural language expressions[8]. An supervised three-component framework to expand some pseudo contexts from web is proposed by Zhao, which can obtain more useful context information to help disambiguate a collocation's polarity [9]. A set of subjective expressions to annotate the contextual polarity in the MPQA Corpus is used by Wilson[10].

III. THE PROPOSED APPROACH

A. Overview

The motivation of our approach is to disambiguate polarity-ambiguous words making full use of sentiment exception of aspects. First, a mutual bootstrapping algorithm is designed to automatically extract polarity words and aspects, utilizing relationships among aspects, polarity words and syntactic patterns. Each time, two sets of the three is fixed to constantly update the third set among aspect, polarity word and syntactic patterns. Secondly, infer the sentiment expectation of an aspect utilizing the relations between aspects and polarity-ambiguous words in annotated reviews. At the same time, more polarity-ambiguous words can be retrieved utilizing the relations between aspects and annotated reviews. Then construct two lexicons one is aspects with sentiment expectation and another is polarity-ambiguous words with prior polarity. Finally, the sentiment of polarity-ambiguous words in context can be decided collaboratively by sentiment expectation of the aspect modified by the polarity-ambiguous words and prior polarity of polarity-ambiguous words.

B. Mutual Bootstrapping Algorithm

Input: corpus S with sentence tag SO(s); seeds polarity-ambiguous words(PAWs) set W1 and score S2(wi) wi ∈ W1; syntactic patterns set R1 and score S3(Ri), Ri ∈ R1; part-of-speech patterns set P; iteration number M and candidate selection number k1, k2.

Output: Dic_aspect and Dic_word.

Initialize aspects set Dic_aspect=Ø; Initialize Dic_PAWs =

W1; Initialize syntactic patterns R_syntactic=R1.

Tokenize each sentence s ∈ C with lexical analysis using ICTCLAS and syntax analysis using ltp-cloud.

3.for m = 1 . . . M do

4.Extract new aspects to Dic_aspect from corpus S as follows

1) For any word w in sentence s ∈S, if w ∈Dic_PAWs.
Within the window of q words previous or behind to w:

If there is a noun phrase along with w meet patterns in P,

put noun phrase into Candi_aspect;

If there is a noun phrase along with w meet the patterns in

R_syntactic, put noun phrase into Candi_aspect;

2) Use aspect pruning strategies to filter out error aspects in Candi_aspect.

3) Update aspect score S1(Ai), where Ai ∈Candi_aspect based on (1), select the top k1 aspects to A1.

4) Infer the sentiment expectation of an aspect a ∈A1 as follows:

For each ap ∈ A1, if w ∈ Dic_PAWs and (ap,w) is a snippet in sentence s ∈ S.

If SO(s)=1,SO(w)=1,SO(ap)=1,Freq(ap+)++;

If SO(s)=0,SO(w)=0,SO(ap)=1,Freq(ap+)++;

If SO(s)=1,SO(w)=0,SO(ap)=0,Freq(ap-)++;

If SO(s)=0,SO(w)=1,SO(ap)=0,Freq(ap-)++;

5) If Freq(ap+)<Freq(ap-),SO(ap)=0,else ap=1.Add aspect ap with sentiment expectation into Dic_aspectand remove repeated aspects.

5.Extract new polarity-ambiguous words to Dic_word from

corpus S as follows.

1) For any nouns phrase ap in sentence s ∈S, if ap ∈ Dic_aspect.
Within the window of q words previous or behind to ap:

if there is a word along with ap is adjective or verb, put

w into Candi_word;

if there is a word along with ap meet the patterns in

R_syntactic, put w into Candi_word;

2) Use polarity word pruning strategies to filter out error polarity words in Candi_word.

3) Update polarity score S2(Wi), where Wi ∈Candi_word based on (2),select the top k2 words toW2.

4) Obtain polarity-ambiguous words from W2 as follows.

For each word w ∈ W2, if ap ∈ Dic_aspect and (ap,w) is a snippet in sentience s ∈ S, SO(w) is prior polarity of w in basic polarity lexicon.

If SO(a)=0, SO(s)=1, SO(w)=0 then w is polarity-ambiguous word;

If SO(s)=0, SO(a)=0,SO(w)=1 then w is polarity-ambiguous word;

5) Add polarity-ambiguous words to Dic_word and remove repeated words.

6.Extract new syntactic patterns to R_syntactic as follows:

1) If w ∈Dic_word, ap ∈Dic_aspect, (ap,w) is a snippet in sentience s ∈S, extract syntactic pattern of w and ap to R2.

2) Update pattern scoreS3(Rj), Where Rj ∈ R2,based on(3), and select top k3 patterns to the pattern set R_syntactic and remove repeated syntactic patterns.

7.end for.

8.return the lexicon of Dic_aspect and Dic_word.

Using the above algorithm a number of PAWs and aspects in different domains can be abstracted. After the iterative process, incorrect PAWs and aspects may be involved in.So we'd better rectify the result manually.

Initiation

- The mutual bootstrapping begins with a seed polarity-ambiguous word set W1. W1 is grouped into two sets: positive-like adjectives (Pa) and negative-like adjectives (Na):Pa and Na are prior polarity of sentiment words in lexicon out of context, but the real positive or negative polarity in context will be evoked when they co-occur with target aspects.

  Pa={高|high,长|long,重|heavy,厚|thick,深 |deep,多|many}

  Na={低|low,短|short,轻|light,薄|thin,浅|shallow,少|less }

- syntactic patterns set R1

These syntactic patterns in R1 are representative and manually selected from syntactic relations between aspects and polarity words using parsing machine. Score ranges from 1 to 10.

(1)NN<===>amod<===>JJ;(2)NN<===>nsubj<===>JJ;

(3)NN<===>dobj<===>VB;

(4)NN<===>conj_and<===>NN<===>amod<===>JJ;

(5)NN<====>dobj<====>VB <====>conj_and<===>JJ;

- part-of-speech patterns set P

These part-of-speech patterns are made up of two parts, one is the sequence of part-of-speech patterns set. We use these patterns to locate exactly target noun which is the component of an aspect modified by polarity word. Another part is the noun or verb phrase patterns set[8].We use these patterns and the target nouns to find noun phrases or verb phrases which are candidate aspects.

As we all know an aspect consists of n characters w=c1,c2,⋯⋯cn, including nouns or verb. First a part-of-speech parser is applied to the reviews[14].The noun is located as the target nouns if the tags of its surrounding consecutive words conform to any of the patterns in Fig.1 part A.

Then consecutive words including target nouns are extracted as candidate aspects from the review if their tags conform to any of the patterns in Fig.1 part B.

Part A :(1)NN+RB+JJ

(2)NN+JJ

(3)NN+VB+"的|of"+JJ

(4)JJ+NN

Fig. 1. part-of-speech patterns set P

[a] NN means nouns, RB means adverb, JJ means adjective.

- Formula in above Algorithm:

$$S1(A_i) = con(A_i) \times \log_2(Freq(A_i)) \times \sum_{R_k \in R} S3(R_k) \quad (1)$$

$$S2(W_j) = con(W_j) \times \log_2(Freq(W_j)) \times \sum_{R_k \in R} S3(R_k) \quad (2)$$

$$S3(R_k) = S1(A_i) \times S2(W_j) \quad (3)$$

$S1(A_i)$ is the score of each aspect，R is syntactic patterns set，$Con(A_i)$ (5)is the PMI of each aspect with aspect set A1, $Freq(A_i)$is the frequency of aspect A_i in corpus，$S3(R_k)$ the score of syntactic patterns using which aspect is extracted. $Con(w_j)$ (7) is the PMI of each polarity with PAWs set W1. $Freq(w_j)$is the frequency of w_j in corpus. $S3(R_k)$is the score of syntactic patterns using which polarity word is extracted.

C. Sentiment Expectation of Aspects

1) Aspect pruning

Not all aspects extracted by syntactic patterns and part-of-speech patterns are useful or genuine aspects. There are also some uninteresting and redundant ones. Aspects pruning aims to remove these incorrect aspects. We use three types of pruning strategies [13].

a) word frequency filtrate:Filter out aspects with low frequency.

b) p-support (pure support): For each aspect t, assuming that the number of sentence including t is s and in these sentence the number of t alone as an aspect rather than a subset of another aspect phrase is k. So we define support=k/s, if the value of support is 0.5, then we recognize t is not a genuine aspect.

c) aspect filtrate based on PMI:

$$PMI(a,b) = N_{ab} / N_{\alpha} \times N_b \quad (4)$$

$$Con(a_i) = \sum_{aj \in Dic} PMI(a_i, a_j) \quad (5)$$

Here N_{ab} is the text number including aspects a and b. N_a is the text number only including aspect a. N_b is the text number only including aspect b. Dic is a set consists of 10 manually selected relevant aspects and product for each product domain as aspect set A.

2) Infer Sentiment Expectation of Aspects:

Sentiment expectation (SE) of aspects is divided into two categories: positive expectation and negative expectation. For a positive expectation aspect, people usually expect the thing referred to by the aspect to be bigger, higher or happen frequently. On the contrary, for a negative expectation noun , people usually expect the thing referred to by the aspect to be smaller, lower or don' t happen . For example, "成本 cheng-ben|cost " is a negative expectation aspect. However, "质量 zhi-liang|quality " is a positive expectation aspect, as most people in most cases expect that their salaries become high.

The So of most snippets consists of aspects and polarity-ambiguous words can be determined by the sentiment expectation of aspects and prior polarity of the polarity-ambiguous words. If the polarity-ambiguous word has the same polarity as the SE of aspect, then the snippet has positive sentiment: if the polarity-ambiguous word has the opposite polarity to the SE of aspect, the snippet has negative sentiment. For example, snippet "成本高|high cost" has negative polarity, because the polarity-ambiguous word "高|high" has positive prior polarity opposite to the SE of aspect "成本|cost" which has negative polarity. While snippet "质量高|high quality" has positive polarity, because the polarity-ambiguous word "高|high" has positive prior polarity the same as the SE of aspect "成本|cost" which has positive polarity.

Relations among aspects, polarity-ambiguous words and snippets can be expressed by the Logic Truth Table below in Table 1.The value 1 on behalf of positive polarity and 0 on behalf of negative polarity.

TABLE I. LOGIC TRUTH TABLE

S(a)	S(w)	S(col)
1	1	1
0	0	1
0	1	0
1	0	0

Here S(a) is the SE of aspects, S(w) is the polarity of the PAWs, S(col) is the polarity of snippets. Combining the Logic Truth Table with polarity relations among aspects, PAWs and snippets, we can deduce formula as follows.

$$S^+(col)=S^+(a)\odot S^+(w)$$
$$S^+(col)=S^-(a)\odot S^-(w)$$
$$S^-(col)=S^-(a)\odot S^+(w)$$
$$S^-(col)=S^+(a)\odot S^-(w)$$

In order to derive the SE of aspects, we transform the above formulas as the following ones, which also meet the Logic Truth Table.

$$S^+(a)=S^+(col)\odot S^+(w)$$
$$S^+(a)=S^-(col)\odot S^-(w)$$
$$S^-(a)=S^+(col)\odot S^-(w)$$
$$S^-(a)=S^-(col)\odot S^+(w)$$

In the above formulas $\odot$ means Not Exclusive Or, $S^+(w)$ means the positive category of PAWs, $S^-(w)$ means the negative category of PAWs; $S^+(a)$ means the positive sentiment expectation of aspects, $S^-(a)$ means the negative sentiment expectation of aspects; $S^+(col)$ means the positive category of snippets, $S^-(col)$ means the negative category of snippets. The polarity of snippets can be obtained by the annotated reviews. In this paper, we hold the assumption that all snippets in the same review have the same polarity as the review's .And the prior polarity of PAWs is fixed in PAWs lexicon.

Considering that one aspect may appear in different snippets and modified by different PAWs. So each aspect may have different SE in different snippets co-occuring with different PAWs. The way to accurately obtain the SE of aspects is based on statistical method. First, we extract snippets consisting of aspects and PAWs using the process in Algorithm from annotated reviews. Secondly, we compute the SE of aspect in each snippets using the formulas in Fig.2 and count the frequency of positive SE Freq(i+)and negative SE Freq(i-)of each aspect. Thirdly, the real SE of aspects can be calculated like this, if Freq(i+) less than Freq(i-), the SE of aspect is negative, otherwise the the SE of aspect is positive.

D. Obtain Polarity-Ambiguous Words

The extraction of polarity word use the same syntactic patterns set R_syntactic as aspects. while the part-of-speech patterns are just adjectives and verbs. We consider the adjectives surrounding aspects are candidate polarity word, but only the emotion verbs surrounding aspects are candidate polarity. The polarity of verbs can derived from the basic polarity lexicon.

1) Polarity word pruning based on PMI:

$$PMI(a,b) = N_{ab} / N_a \times N_b \quad (6)$$

$$Con(w_i) = \sum_{wj \in Dic} PMI(w_i, w_j) \quad (7)$$

Here N_{ab} is the text number including polarity word a and polarity word b. N_a is the text number only including polarity word a. N_b is the text number only including polarity word b. Dic is set W1 used in Algorithm which contain 12 frequently used PAWs.

2) Infer Polarity-ambiguous words(PAWs)

Independent polarity word can accurately express the sentiment individually, such as "happy", "sad". While the sentiment of PAWs in context should rely on the SE of aspects, such as "高|high","长|long". So this is an indication of how to distinguish polarity words and PAWs. When inferring a polarity word we'd better take the snippet and aspect into consideration. If the polarity of snippet decided by the polarity individually, on the other hand the polarity of snippet agree with the polarity word, then we define the polarity word is an independent polarity word. If the polarity of snippet opposite to the polarity word, then we define the polarity word as PAWs. For example snippet"价格合理"means favorable price which has positive polarity, the same as "favorable", so "favorable" is an independent polarity word. While snippet"价格高"means high

price, which has negative polarity, opposite to "高|high", so "高|high" is a polarity-ambiguous word.

Special cases should be taken into consideration, when aspects are positive, the inference is not established. For example, snippet" 质 量 好 "means good quality which has positive polarity the same as "good". As we all know "good" is an independent polarity word. While the snippet"质量高"means high quality, which also has positive polarity the same as "高|high". But as we all know "高|high" is a polarity-ambiguous word. So the above assumption is valid when SE of the aspect is negative in snippets .This also prove the necessity of construct aspects lexicon with SE. Using this method we can find more PAWs and the polarity is its prior polarity.

IV. Experiments and Results

A. Sentiment Analysis at Sentence Level

In order to test the performance of the PAWs lexicon and aspect lexicon constructed in this paper, we did some experiments.

1) Data and Preproccess

We collected data from popular forum sites it168, JingDong, DataTang. Reviews in different domains such as book, computer and so on are grabbed. In each domain we manually annotated 3000 positive reviews and 3000 negative reviews as train corpus, 500 positive reviews and 500 negative reviews as test corpus on sentiment analysis at sentence level. In order to concentrate on the disambiguation of PAWs, and reduce the noise introduced by the parser, we extracted sentences for test corpus containing at least one adjective and one aspect in a sentience.

The reviews were automatically word segmented and POS-tagged using the open software ICTCLAS [14].The reviews were also automatically syntactic analysised using software ltp-cloud[15].

2) Evaluation Metrics

Instead of using accuracy, we use precision (P), recall (R) and F1-value (F1) to measure the performance of sentiment analysis at sentence level. We establish the mixed matrix as shown in Table3. Mixed matrix is special to each category and it count the classification of each sentience.

TABLE II. TABLE2:MIXED MATRIX

	classified as positive	**classified as negative**
real positive	TP	FP
real negative	FN	TN

$$Recall = \frac{TP}{TP + FP} \quad (5)$$

$$Precision = \frac{TP}{TP + FN} \quad (6)$$

$$F1 - value = \frac{2 \times Recall \times Precision}{Recall + Precision} \quad (7)$$

3) Methods

Our goal is not to propose a new method, but instead to test the performance of aspect and PAWs lexicon we constructed. We adopted the same algorithm with Wan (2008)[16], and we not only use Sentiment-HowNet but also NTUSD as basic polarity lexicon. But in our experiment, Intensifier_Dic didn't use.

Algorithm Compute_SO:

1) Tokenize each sentence s ∈S into word set Ws and PAWs;

2) For any word w in a sentence s ∈S, compute its value SO(w) as follows:

1) if w∈PAWs, compute SO(w)

a) In baseline1 method only use the PAWs lexicon

If SO(w)=1, SO(w)=Dy_PosValue;

If SO(w)=0, SO(w)= Dy_NegValue

b) In baseline2 method ,use the PAWs lexicon and aspect lexicon constructed by Zhou[18]

Within the window of q words previous or behind to w, if there is a term a∈aspect lexicon.

If SO(w)=1,SO(a)=1,SO(w)=Dy_PosValue;

If SO(w)=1,SO(a)=0,SO(w)=Dy_NegValue;

If SO(w)=0,SO(a)=1,SO(w)=Dy_NegValue;

If SO(w)=0,SO(a)=0,SO(w)=Dy_PosValue;

c) In our method , use the PAWs lexicon and aspect lexicon constructed by this paper.

Within the window of q words previous or behind to w, if there is a term a∈aspect lexicon A.

If SO(w)=1,SO(a)=1,SO(w)=Dy_PosValue;

If SO(w)=1,SO(a)=0,SO(w)=Dy_NegValue;

If SO(w)=0,SO(a)=1,SO(w)=Dy_NegValue;

If SO(w)=0,SO(a)=0,SO(w)=Dy_PosValue;

2)If w∈Positive_Dict, SO(w)=PosValue;

3)If w∈Negative_Dict, SO(w)=NegValue;

4)Otherwise, SO(w)=0;

5)Within the window of q words previous to w, if there is a term w'∈Negation_Dict.

SO(w)= – SO(w);

3. $S(s) = \sum_{w \in Ws} SO(w)$

a) Baseline1: Not considering the context, assign all positive-like adjectives as positive, and all negative-like adjectives as negative.

b) Baseline2: Use aspect and PAWs lexicon constructed by Zhou[17]

c) Our method: Use aspect and PAWs lexicon lexicon constructed by this paper.

B. Result:

The performance of sentiment classification of product reviews in two domains which is book and computer was significantly improved. In each domain we use 500 positive reviews and 500 negative reviews, The result is shown in Table3:

TABLE III. THE EXPERIMENTAL RESULTS AT SENTIENCE LEVEL

		Baseline1		Baseline2		Our method	
		B	C	B	C	B	C
Pos	Pre	0.6740	0.7165	0.6899	0.7401	0.7123	0.7540
	Rec	0.7650	0.8310	0.7902	0.8645	0.8050	0.8990
	F1	0.7166	0.7696	0.7366	0.7974	0.7558	0.8201
Neg	Pre	0.7283	0.7797	0.7543	0.8284	0.7759	0.8687
	Rec	0.6301	0.6453	0.6450	0.6829	0.6750	0.6950
	F1	0.6756	0.7062	0.6954	0.7487	0.7219	0.7722
Aver		0.6916	0.7379	0.7160	0.7731	0.7539	0.7961

[b.] B means book,C means computer， aver is the average of F1

Adding the disambiguation of PAWs, our method obviously outperforms the baseline1, especially in computer reviews. People usually use more PAWs in smart devices reviews. But in book domain the classification result is lower than in other domain, because in book reviews less PAWs and negative aspects are used. In some book reviews, there are words just describing the content of books which disturbs classify of reviews. And the classification result in negative reviews is lower than positive reviews. This is because in positive reviews people usually use more independent polarity words to express their emotion. While in negative reviews people tend to describe the property of products more frequently rather than express their emotion, so less independent polarity words are used. Our method also outperforms the baseline2 just a little bit, which prove that our method can recognize more PAWs and aspects with SE, though the quantity is not large.

V. CONCLUSION

This paper presents a mutual bootstrapping algorithm to construct aspect lexicon and polarity-ambiguous lexicon in order to disambiguate the polarity-ambiguous words in the context. When a polarity-ambiguous word appears in a sentence, firstly extract the aspect around the PAWs , then find it's SE in aspect lexicon and find the polarity of the PAWs in Polarity-Ambiguous Word lexicon, finally compute the real polarity of the PAWs in sentence using the SE of aspect and prior polarity of the PAWs. For the sentiment analysis at sentence level, our method achieves promising result that is significantly better than baseline and automatically extract more polarity-ambiguous words rather than only 14 polarity words used in baseline. On the other hand compared to others manual extract methods, our method automatically extract aspects and polarity words which reduce the manually work and achieve obvious improvement in performance. This validates the effectiveness of our approach.

There leaves room for improvement. In this paper method of extracting the aspects and polarity words always generate some noises, so find out new methods to reduce noises is our future work. The mutual bootstrapping algorithm in this paper need annotated reviews which bring in manual operation. SO discover efficient unsupervised method without manual operation in inferring the SE of aspects and construct aspect and PAWs lexicon is the future work.

REFERENCES

[1] Bo Pang and Lillian Lee, A Sentimental Education: Sentiment Analysis Using Subjectivity Summarization Based on Minimum Cuts, Proceedings of ACL 2004.

[2] Peter Turney and MichaelL.Littman.2003. Measuring praise and criticism: Inference of semantic orientation from association. ACM TOIS, 21(4):315–346.

[3] Ding, X., Liu, B. and Yu, P. 2008. A holistic lexicon-based approach to opinion mining. In Proceedings of WSDM'08.

[4] Yasuhiro Suzuki, Hiroya Takamura, and Manabu Okumura. 2006. Application of semi-supervised learning to evaluative expression classification. In Computational Linguistics and Intelligent Text Processing, pages 502–513.

[5] Wu,Y and Wen, M. 2010. Disambiguating Dynamic Sentiment Ambiguous Adjectives. In Proceedings of COLING2010.

[6] Wen, M and Wu,Y.2011. Mining the Sentiment Expectation of Nouns Using Bootstrapping Method. Proceedings of the 5th IJCONL2011, 1423–1427.

[7] Hu, Minqing and Bing Liu. 2004. Mining and summarizing customer reviews. In Proceedings of ACM SIGKDD Conference on Knowledge Discovery and Data Mining 2004 (KDD-2004), pages 168–177, Seattle, Washington.

[8] Ding, X, Liu, B. and Yu, P. 2008. A holistic lexicon-based approach to opinion mining. In Proceedings of WSDM'08.

[9] Chunliang LI, Yanhui Zhu, Yeqiang Xu.Product reviews in Chinese word extraction method of research[J].Computer Engineering,2011,37(12):26-32

[10] Yanyan Zhao, Bing Qin and Ting Liu. Collocation Polarity Disambiguation Using Web-based Pseudo Contexts. In Proceedings of the 2012 Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL 2012). 2012.07. Jeju, Republic of Korea. (long, oral)

[11] Theresa Wilson, Janyce Wiebe, Paul Hoffmann. Recognizing Contextual Polarity: An Exploration of Features for Phrase-Level. Computer Linguistics,25(3),399-433

[12] Turney P. Thumbs up or thumbs down? semantic orientation applied to unsupervised classification of reviews[C]Proceedings of ACL'02. Pennsylvania: ACL, 2002: 417-424.

[13] Liu Hongyu,Zhao Yanyan,Qin B,LIu T. Aspect Extract and Analysis[J]. Journal of Chinese information.2010,24(1):84-89.

[14] ICTParser: http://nlp.ict.ac.cn/demo/ictparser/index.php.

[15] Harbin industrial university language technology platform.IR: http://ir.hit.edu.cn/demo/ltp/.

[16] Wan, X. 2008.Using Bilingual Knowledge and Ensemble Techniques for Unsupervised Chinese Sentiment Analysis. Proceedings of EMNLP'08.

[17] Zhou C, Sentiment analysis technology research for Chinese Web comments [D]. National University of Defense Technology.2011.

GPS-Based Daily Context Recognition for Lifelog Generation Using Smartphone

Go Tanaka
Graduate School of Informatics,
Shizuoka University
Hamamatsu, Japan

Masaya Okada
Graduate School of Informatics,
Shizuoka University
Hamamatsu, Japan

Hiroshi Mineno
Graduate School of Informatics,
Shizuoka University
Hamamatsu, Japan

***Abstract*—Mobile devices are becoming increasingly more sophisticated with their many diverse and powerful sensors, such as GPS, acceleration, and gyroscope sensors. They provide numerous services for supporting daily human life and are now being studied as a tool to reduce the worldwide increase of lifestyle-related diseases. This paper describes a method for recognizing the contexts of daily human life by recording a lifelog based on a person's location. The proposed method can distinguish and recognize several contexts at the same location by extracting features from the GPS data transmitted from smartphones. The GPS data are then used to generate classification models by machine learning. Five classification models were generated: a mobile or stationary recognition model, a transportation recognition model, and three daily context recognition models. In addition, optimal learning algorithms for machine learning were determined. The experimental results show that this method is highly accurate. As examples, the F-measure of the daily context recognition was approximately 0.954 overall at a tavern and approximately 0.920 overall at a university[1].**

Keywords—component; Lifelog; machine learning; GPS; healthcare

I. INTRODUCTION

The use of mobile devices such as smartphones and tablets has become more widespread and sophisticated. These devices contain many diverse and powerful sensors, such as GPS sensors, acceleration sensors, and gyroscope sensors. Since the sensors are very small and lightweight, they can collect various types of personal data without inconveniencing the user. Mining the collected data helps us to learn many details about the daily lives of people. The knowledge gained from data mining can be applied to various individual life support services such as healthcare services and online-to-offline (O2O) services. Individuals can also use the knowledge obtained from a lifelog recorded by the model devices. The lifelog, which captures all or a large portion of one's life, can be referenced to learn about one's own experiences and lifestyle. Services related to lifelogs have recently been attracting attention [1]–[3]. High-quality lifelogs can speed up the development of essential services, especially in healthcare. This is an important goal due to the rapid increase in the number of people who die from lifestyle-related diseases such as cancer, heart disease, and strokes. The World Health Organization (WHO) refers to lifestyle-related diseases as noncommunicable diseases (NCDs) and states that 36 million people, or 63% of the 57 million global deaths, died from NCDs in 2008 [4]. The WHO further estimates that the total number of annual NCD deaths is projected to reach 55 million by 2030. In Japan, NCDs have become the primary cause of death. Therefore, people should examine their own daily lives to improve their lifestyles.

In an extension of a previous study [5], this paper proposes a method for daily context recognition. This method generates high-quality lifelogs by using only a GPS sensor. Because locations and activities are important elements for a daily lifelog, this method recognizes the user's location and activity as contexts. It also helps to distinguish and recognize several contexts that are considered to appear at the same location.

The proposed method measures several variables captured by a commercial mobile phone for daily context recognition of a person's activity. The user's activities captured by the mobile phone show their habits over time. Therefore, the user can look back at his or her daily life in more detail by routinely carrying a smartphone. In addition, it becomes possible to create and provide services in accordance with the user's location and situation. The intended result is an improved lifestyle for the prevention of lifestyle-related diseases.

II. RELATED WORK

Lifelog generation using human activity recognition technologies has recently gained attention as a research topic. Research on healthcare for lifestyle-related diseases has also been increasing. Much of the research focuses on monitoring and recognizing human activity to give feedback to the user. Results of some of the research on human activity recognition and healthcare related to lifelogs are described below.

A. Human Activity Recognition

In one of the earliest studies on human activity recognition, Lara and Labrador [6] used three sensing devices—GPS, accelerometers, and vital signs—and created decision tree models to recognize three basic physical activities—walking, running, and sitting—by using the C4.5 classification algorithm. Gomes et al. [7] developed a mobile activity recognition system (MARS) that learns the classification model onboard

[1] G. Tanaka is with the Graduate School of Informatics, Shizuoka University, Japan

M. Okada is with the Graduate School of Informatics, Shizuoka University, Japan

H. Mineno is with the Graduate School of Informatics, concurrently with the Graduate School of Science and Technology, the Research Institute of Green Science and Technology, Shizuoka University, Japan

the mobile device itself through ubiquitous data stream mining in an incremental manner. Using the naive Bayes classification algorithm along with acceleration data, they created personal models to recognize five physical activities—walking, running, standing still, driving, and climbing stairs. Kwapisz et al. [8] used accelerometers to recognize daily physical activities. They compared three classification algorithms—the J48 decision tree, logistic regression, and the multilayer perceptron—and found that the multilayer perceptron performed the best overall for recognizing six physical activities and actions (walking, jogging, climbing stairs, going down stairs, sitting, and standing). Hattori et al. [9] developed the ALKAN system. ALKAN is a server-client system that gathers a large number of "missions" by using mobile sensor devices with accelerometers. They recognized eight physical activities—eating, cycling, riding in a car, sitting, standing, sitting in a train, standing in a train, and walking. For machine learning training, they used four classification algorithms: the recursive partitioning tree, naive Bayes classifier, nearest neighbor classification, and support vector machine (SVM). Many other researchers have also used machine learning for human activity recognition, and some of them have used accelerometer data [10]–[18].

All of the above studies recognize only basic human physical activities or attitudes; none recognize the purpose for an activity in a person's daily life. Therefore, these methods do not obtain the information for generating lifelogs, and users cannot look back at the contexts of their daily lives to improve their lifestyles. Furthermore, some problems must be considered when using acceleration sensors. First, the accuracy of the recognition when using acceleration sensors depends on the mounting position of the sensor device. The mounting position of wearable sensors is especially important when the user's mobile phone contains an accelerometer. Second, it is impossible to gather accurate acceleration sensing data when the user is actively using his or her phone, because too much noise is generated while touching the smartphone.

The method presented in this paper is a novel method that can gather accurate sensing data regardless of the wearable position and operation, and can accurately recognize specific contexts of daily life.

B. Healthcare

In the healthcare research field, many participants use wearable sensor devices or smartphones with acceleration sensors, GPS, and so on, for monitoring and managing the user's lifestyle and personal medical information [19]–[29]. Suzuki et al. [19] proposed improving lifestyles from the perspective of a user's meal choice. They analyzed the user's situation, biological information, and lifestyle (e.g., budget, workload, and spare time) and then provided advice on purchasing appropriate foods. Yan et al. [20] used body sensors for monitoring the movement of elderly people. They proposed a mixed positioning algorithm to determine the location of an elderly person to determine that person's activities and to make decisions about his/her health status. Khan et al. [21] used an accelerometer in a smartphone to recognize a person's daily physical activities in order to suggest minor behavioral modifications to increase the energy expenditure in one's daily routine.

Although most of the above studies use a wearable device, the focus is almost entirely on treating diseases and health problems, and there is little focus on actually preventing them. According to the WHO, the highest probability of dying from an NCD is between the ages of 30 and 70 in many locations around the globe, and the accrual of improper lifestyles from a young age can be thought of as one of the reasons for a later NCD; that is, by the time a person is old, it may be too late. Therefore, it is necessary for people to monitor and improve their own daily lifestyles from a young age to prevent later lifestyle-related diseases.

III. Daily Context Recognition

The method proposed in this paper has the following requirements:

1) Recognizing not only basic physical activities but also specific contexts related to daily life.

2) Gathering accurate data regardless of the wearable position and operation of the sensor device.

3) Reducing the installation cost so that young people can use it.

4) Preventing lifestyle-related diseases by improving the user's lifestyle.

The proposed method estimates the user's location and his or her contexts at any given moment. For example, several contexts can occur at one given location, such as a shopping mall. This method can distinguish between shopping, eating a meal, and seeing a movie at the shopping mall. Daily activities (with the exception of home activities, which were the focus of a previous paper [30]) are estimated in this paper.

The contexts for lifelog generation are estimated by data mining using sensor data. This paper uses a GPS sensor in a smartphone because such sensor data do not depend on the user's position or operation. Moreover, because individual differences rarely appear, it is possible to generate a generic model. The user needs only a smartphone, and so the installation cost is low. All of this is intended to generate lifelogs for the prevention of lifestyle-related diseases.

Figure 1 shows a flowchart for recognizing daily contexts. Most recognition processors use machine learning. The Weka [31] data mining software was used for the machine learning in this work. Details of the proposed method are described in the following.

A. Feature Extraction

The feature extraction processor extracts five features for machine learning: speed, variance, weather, time zone, and day of the week. First, the GPS obtains the latitude, longitude, and time. The speed is calculated on the basis of the distance between two location points, which are determined by latitude (*lat*) and longitude (*lon*) values, and then this value is divided by the difference between the times of data collection. The variance of location is calculated in (1). In this paper, the number of longitudinal data *n* is 5.

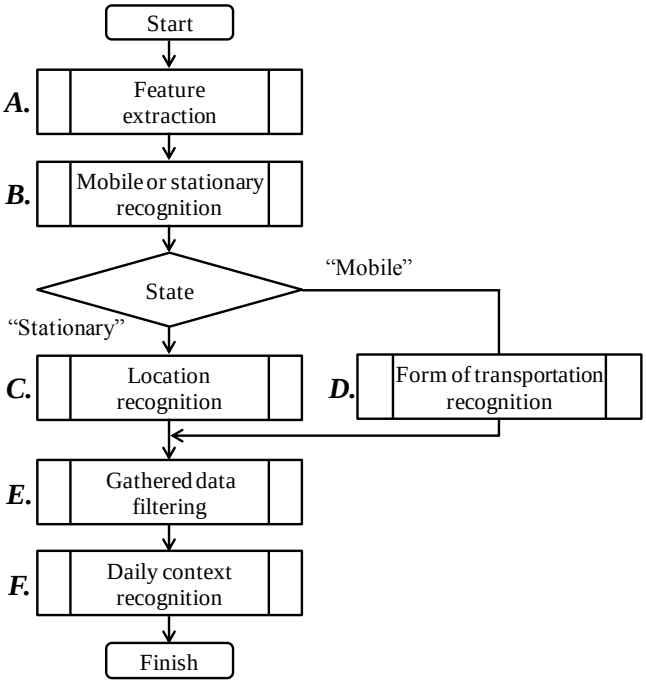


Fig. 1. Flowchart of user's daily context recognition using only a GPS sensor

$$\sqrt{\left\{\frac{1}{n}\sum_{i=1}^{n}\left(\overline{lat}-lat_i\right)^2\right\}^2+\left\{\frac{1}{n}\sum_{i=1}^{n}\left(\overline{lon}-lon_i\right)^2\right\}^2} \quad (1)$$

Weather is obtained from a Meteorological Agency site according to the given time. The time zone is defined as late_night (0:00 to 2:00), early_morning (2:00 to 6:00), forenoon (6:00 to 10:00), noon (10:00 to 14:00), afternoon (14:00 to 18:00), night (18:00 to 22:00), and late_night (22:00 to 24:00).

B. Mobile or Stationary Recognition

The next step is to recognize whether a user is mobile or stationary in order to recognize their location. The mobile or stationary condition needs to be simply and very accurately recognized because this process is the first step in recognizing daily contexts. Machine learning is used for recognizing whether the user is mobile or stationary. The explanatory variables for machine learning are the speed and the variance of the location. The user is considered to be mobile if the speed or the variance is greater, and vice versa.

C. Location Recognition

This processor recognizes a stationary location by recognizing whether the person is mobile. By registering the information of locations in a database well in advance, it is possible to easily estimate the user's location when they are stationary at a registered position. If the user is stationary at an unregistered position, it is possible to obtain information from nearby facilities by a place search API (a service that returns information about places by HTTP requests).

D. Form of Transportation Recognition

This processor recognizes the type of transportation when the person is mobile. The form of transportation is used to determine the number of non-exercise activity thermogenesis (NEAT) calories and is one of the explanatory variables in machine learning. Since NEAT includes all daily life activities, it is possible to prevent lifestyle-related diseases by promoting non-exercise daily activities. The type of transportation used is an explanatory variable because people change the form of transportation on the basis of the purpose and situation at their given location. By also using speed, location variance, and the weather as explanatory variables, it is possible to recognize five forms of transportation: on foot (Walk), riding a scooter (Scooter), driving in a car (Car), taking a train (Train), and taking the Shinkansen (Shinkansen).

E. Data Filtering

The GPS data collected by a smartphone are filtered to improve the level of accuracy for recognizing daily contexts. The positioning error is significantly worse indoors with the GPS sensor. Therefore, a lot of generated noise must be removed by filtering. However, since general filtering methods such as the moving average filter and the low-pass filter are used for smoothing longitudinal data, they cannot respond to rapid and large changes in a user's position (e.g., transition from a stopped state while on the Shinkansen). Therefore, this processor removes the noise data by using accuracy values obtained from the GPS sensor of a general OS-based mobile phone. The level of accuracy is given in meters as the error range of the distance. The accuracy is defined as a 68% confidence radius. By appropriately setting a threshold for the accuracy values, it is possible to improve the recognition accuracy by removing the noise.

F. Daily Context Recognition

Finally, this paper recognizes the contexts for the stationary location related to a person's daily life. It is possible to recognize not only basic human physical activities but also a user's location (i.e., daily location) and context (i.e., what the user is doing there) by generating context recognition models. It is possible to determine a person's poor lifestyle habits, improve them, and help to prevent lifestyle-related diseases by recognizing the contexts that account for most of his or her daily life at a specific location. The location where the contexts have been recognized—e.g., a restaurant, university, convenience store, or shopping mall—is registered in the database, and after that, the same model can be used when the user visits a similar location for the first time.

IV. Experiment

This section describes our initial experiments to evaluate the proposed method for recognizing daily human contexts. As mentioned above, because users need to improve their lifestyles from a young age, experiments were conducted based on the daily activities of students. Figure 2 shows the daily activity model of an example student.

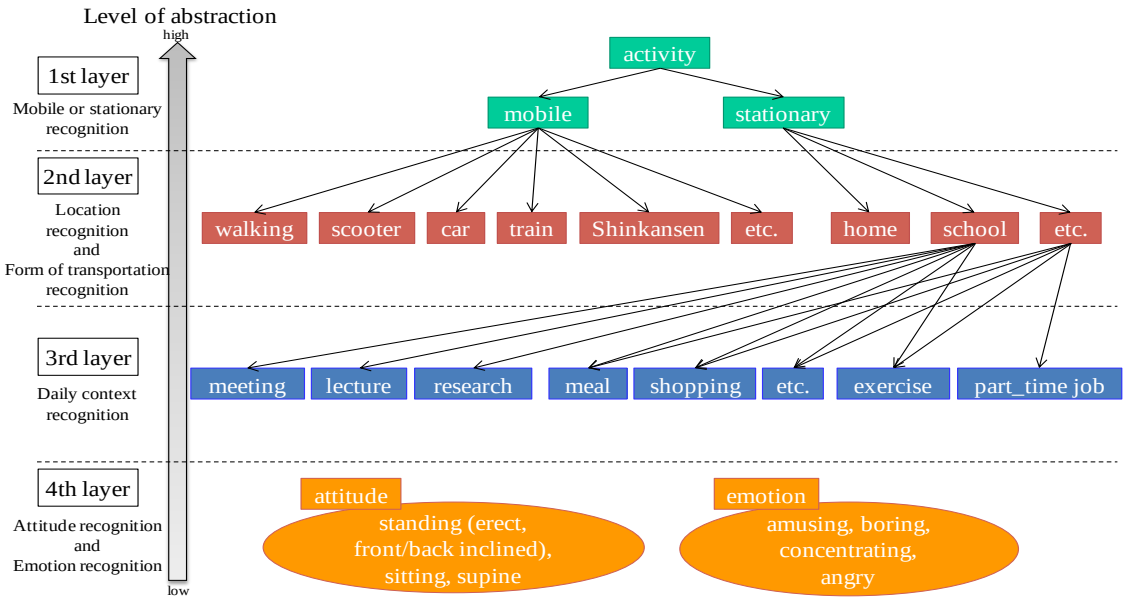


Fig. 2. Daily activity model of a student. Model is divided into four layers referring to different levels of abstraction

The activities of all participants are divided into mobile and stationary. Students walk, ride a scooter, drive a car, board a train, board the Shinkansen, and so on as primary mobile activities while their home, school, etc. are the primary places of stationary activities. Contexts such as meetings, lectures, research, meals, shopping, exercise, and part-time jobs occur at the stationary places. In addition to these activities and contexts, actions such as sitting and standing are recorded at all times. It is clear that people who walk on a regular basis are healthier than people who get into a car or ride a scooter. It is also possible to call attention to people who eat meals at irregular times or who do not eat at all. Moreover, the effect on lifestyle-related diseases depends on the user's context, even when people are stationary at the same location. For example, a student's expenditure of calories increases by working a part-time job at a restaurant, but the calorie intake increases when they eat a meal there. Therefore, it is necessary to distinguish between and recognize the contexts that indicate a difference in calorie consumption. The lifelog generated by recognizing the contexts based on this model could improve the lifestyles of students.

We investigated three details in the experiments: recognition accuracy, validity of the explanatory variables for machine learning, and the optimal machine learning algorithm. The recognition accuracy is evaluated by F-measure F_1, as

$$F_1 = \frac{2 \times recall \times precision}{recall + precision} \quad (2)$$

F_1 is a performance measure widely adopted in the fields of pattern recognition and information retrieval. It is the harmonic means of both recall and precision. Various combinations of explanatory variables were compared in experiments to confirm their validity. Random forest, J48 based on C4.5, SVM, neural network, and Bayesian network algorithms were compared to determine the optimal algorithm for machine learning. Default WEKA functions were employed to ensure a ten-fold cross validation in all experiments. The purpose of machine learning was classification, so we used support vector classification (SVC). Parameters such as cost and gamma were tuned by using a grid search because SVM is an algorithm that considers parameter tuning as the most important operation. The radial basis function (RBF) was used as the kernel function.

A. Experimental Results

The accuracy of mobile or stationary recognition, the form of transportation recognition, and the daily context recognition were evaluated in the initial experiments. The recognized contexts were the locations at a tavern and at a university, both of which are places assumed to be frequented by students in their daily lives.

1) Mobile or stationary recognition

GPS data from five participants were collected for two weeks. Figure 3 shows the accuracies of mobile or stationary recognition by using the random forest algorithm. In the figure, the Gini is defined as the "inequity" of a society's distribution of income, or a measure of "node impurity" in a tree-based classification. TABLE I lists the correctness of the models generated by the five learning algorithms and the accuracies when using the test data.

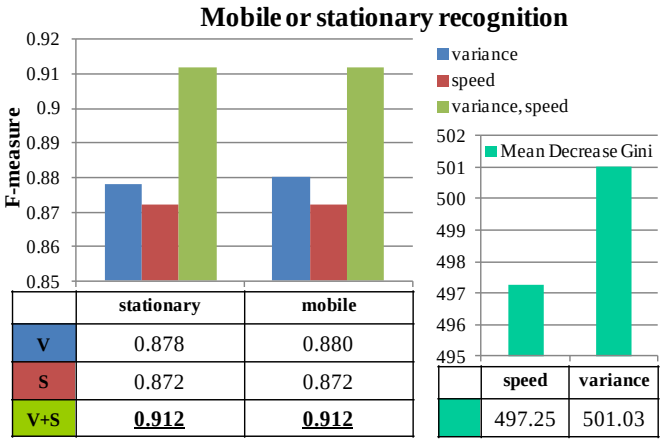


Fig. 3. Accuracies of mobile or stationary recognition by using random forest algorithm for combinations of explanatory variables, and the mean decrease Gini of explanatory variables. Highest values are in bold and underlined

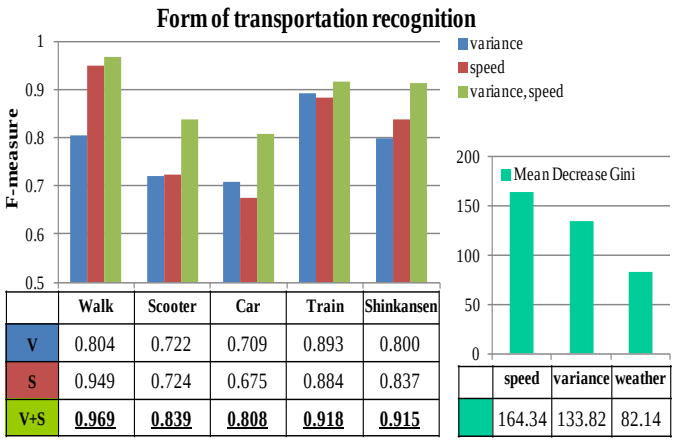


Fig. 4. Differences in accuracies of form of transportation recognition by using random forest algorithm for combinations of explanatory variables, and the mean decrease Gini of explanatory variables

TABLE I. MODEL CORRECTNESS AND ACCURACIES OF TEST DATA FOR MOBILE OR STATIONARY RECOGNITION

Model/ Test	Random Forest	J48	SVM	Neural Network	Bayesian Network
Stationary	*0.912 /* **0.926**	*0.898 /* 0.898	*0.890 /* 0.892	*0.889 /* 0.889	*0.921 /* 0.925
Mobile	*0.912 /* **0.924**	*0.883 /* 0.883	*0.879 /* 0.883	*0.879 /* 0.885	*0.921 /* 0.923
Overall	*0.912 /* **0.925**	*0.891 /* 0.891	*0.884 /* 0.888	*0.884 /* 0.887	*0.921 /* 0.924

Three cases were compared in the experiments: using only the location variance, using only the speed, and using both variables as explanatory variables. As shown in Figure 3, using both variables achieves the highest level of accuracy. It is also clear that the importance of both variables is significantly higher when looking at the mean decrease Gini values. A higher decrease in Gini implies that a particular predictor variable plays a greater role in partitioning the data into the defined classes. TABLE I lists the model correctness generated from a training data set of 2,000 (stationary: 1,000, mobile: 1,000) and the evaluation results when using the training data set and the test data set of 1,000 (stationary: 500, mobile: 500). As shown in this table, all cases achieve high levels of accuracy above 0.880 (F-measure). The random forest algorithm achieves the highest level of accuracy in both activities. An incorrect classification occurs when GPS data with large positioning errors from an indoor location are used in the learning and the data are recognized as "stationary" even if the speed or variance values are large.

2) Form of transportation recognition

The same GPS data collected from the same five participants for two weeks were used in this evaluation. Figure 4 shows the accuracies of the form of transportation recognition using the random forest algorithm. TABLE II lists the correctness of the models generated by the five learning algorithms and the accuracies when using the test data.

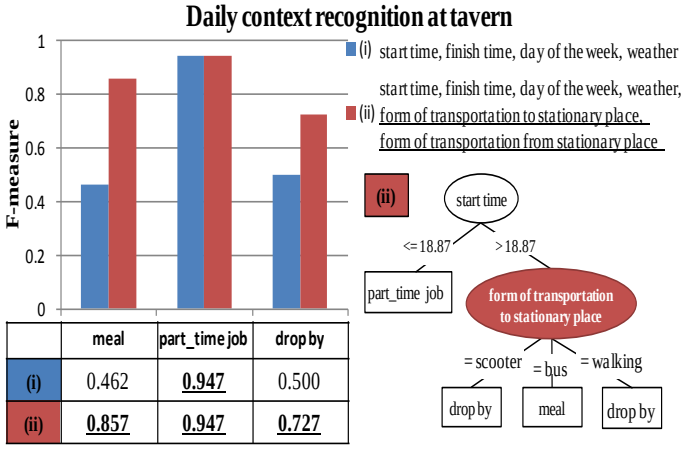
TABLE II. MODEL CORRECTNESS AND ACCURACIES OF TEST DATA FOR FORM OF TRANSPORTATION RECOGNITION

Model/ Test	Random Forest	J48	SVM	Neural Network	Bayesian Network
Walk	*0.969 /* 0.946	*0.959 /* 0.951	*0.962 /* 0.951	*0.917 /* **0.959**	*0.973 /* 0.953
Scooter	*0.839 /* **0.861**	*0.745 /* 0.750	*0.657 /* 0.744	*0.683 /* 0.707	*0.797 /* 0.816
Car	*0.808 /* **0.828**	*0.694 /* 0.698	*0.598 /* 0.686	*0.559 /* 0.638	*0.653 /* 0.622
Train	*0.918 /* **0.901**	*0.881 /* 0.855	*0.779 /* 0.700	*0.833 /* 0.754	*0.822 /* 0.757
Shinkansen	*0.915 /* **0.929**	*0.919 /* 0.893	*0.779 /* 0.813	*0.858 /* 0.863	*0.877 /* 0.863
Overall	*0.890 /* **0.893**	*0.840 /* 0.829	*0.759 /* 0.779	*0.770 /* 0.784	*0.824 /* 0.802

Comparative experiments were performed in cases similar to the experiment for mobile or stationary recognition. However, weather was added to the explanatory variables in all cases because it could affect the user's behavior. As shown in Figure 4, the use of all variables ensures the highest level of accuracy. The mean decrease Gini values show that speed was the most important element for the form of transportation recognition. TABLE II lists the model correctness generated from a training data set of 750 (150 pieces of data per context) and the evaluation results when using the training data set and a test data set of 375 (75 pieces of data per context). As shown in this table, almost all cases are highly accurate, indicating that it is possible to achieve high levels of accuracy in all activities by using the random forest algorithm. The random forest algorithm achieves higher levels of accuracy than the other algorithms for Scooter, Car, and Train because it can generate a strong classifier by group learning of the decision trees. The reason for the lower levels of accuracy for Scooter and Car is assumed to be the frequent changes in car speed stemming from various traffic conditions, such as signal changes and pedestrians, and thus the car speed is often equal to the scooter speed. Therefore, Car is erroneously recognized as Scooter.

3) Daily context recognition at a tavern

GPS data were collected for one participant (one of the authors) for one month. Figure 5 shows the accuracies of the daily context recognition at a tavern by using the J48 algorithm. TABLE III lists the model correctness generated by all five learning algorithms.



	meal	part_time job	drop by
(i)	0.462	**0.947**	0.500
(ii)	**0.857**	**0.947**	**0.727**

Fig. 5. Differences in accuracies of daily context recognition at a tavern by using J48 algorithm for combinations of explanatory variables, and the decision tree

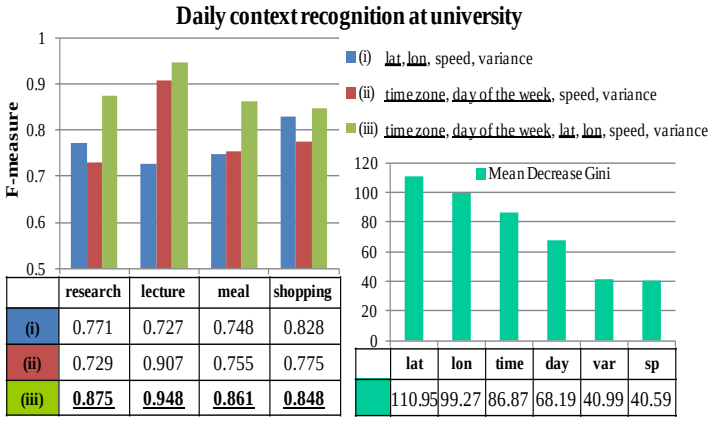
TABLE III. MODEL CORRECTNESS FOR DAILY CONTEXT RECOGNITION AT TAVERN

Model	Random Forest	J48	SVM	Neural Network	Bayesian Network
Meal	*0.667*	*0.857*	*0.600*	***0.923***	***0.923***
Part-time job	*0.952*	*0.947*	*0.952*	***1.000***	*0.947*
Drop by	*0.545*	*0.727*	*0.769*	***0.909***	*0.833*
Overall	*0.763*	*0.863*	*0.806*	***0.954***	*0.910*

Comparative experiments were performed for two cases. One case included the start time, finish time, day of the week and weather as the explanatory variables. The second case added the form of transportation to (from) the stationary place. We confirmed in this experiment that the form of transportation is also effective as an explanatory variable. As indicated in Figure 5, it is possible to improve the levels of accuracy for "meal" and "drop by" by approximately two times and 0.23, respectively. As shown in the decision tree, for the contexts classified as "part-time job" or any of the others at the start time, the form of transportation to the stationary place is used for the classification. This variable achieves higher levels of contextual accuracy for the "meal" and "drop by" variables. This could mean that many young people rarely drive a car or ride a scooter when they are going to have a meal at a tavern because the probability of drinking alcohol there is higher. TABLE III lists the correctness of the model generated from a training data set of 22 (meal: 6, part-time job: 10, drop by: 6). As specified in this table, although the recognition accuracy of "part-time job" is high for all of the algorithms, the overall accuracy of the random forest algorithm is the lowest, unlike in the previous evaluations. We presume that the benefits of group learning are not applicable because the training data set is too small. However, even though the training data set was small, the neural network algorithm achieved high levels of accuracy.

4) Daily context recognition at a university

GPS data were collected for one subject (one of the authors) for two months. Figure 6 shows the accuracies of the daily context recognition at a university by using the random forest algorithm. TABLE IV lists the correctness of the models generated by the five learning algorithms and the accuracies for the test data.



	research	lecture	meal	shopping
(i)	0.771	0.727	0.748	0.828
(ii)	0.729	0.907	0.755	0.775
(iii)	**0.875**	**0.948**	**0.861**	**0.848**

lat	lon	time	day	var	sp
110.95	99.27	86.87	68.19	40.99	40.59

Fig. 6. Differences in accuracies of daily context recognition at a university by using random forest algorithm for combinations of explanatory variables, and the mean decrease Gini of explanatory variables

TABLE IV. MODEL CORRECTNESS AND ACCURACIES OF TEST DATA FOR DAILY CONTEXT RECOGNITION AT UNIVERSITY

Model / Test	Random Forest	J48	SVM	Neural Network	Bayesian Network
Research	*0.875* / **0.912**	*0.831* / 0.861	*0.665* / 0.845	*0.473* / 0.639	*0.666* / 0.763
Lecture	*0.948* / **0.967**	*0.921* / 0.961	*0.874* / 0.944	*0.867* / 0.898	*0.842* / 0.872
Meal	*0.861* / **0.913**	*0.832* / 0.830	*0.762* / 0.887	*0.741* / 0.773	*0.800* / 0.815
Shopping	0.848 / **0.887**	*0.823* / 0.843	*0.729* / 0.834	*0.679* / 0.831	*0.822* / 0.845
Overall	*0.883* / **0.920**	*0.852* / 0.874	*0.757* / 0.878	*0.690* / 0.785	*0.782* / 0.824

Comparative experiments were conducted for three specific cases. Case (i) used the latitude, longitude, speed, and variance as explanatory variables. Case (ii) used the time zone, day of the week, speed, and variance. Case (iii) used all the variables. Case (iii) achieved the highest levels of accuracy. The most effective variables are latitude and longitude and the next most effective variables are time zone and day of the week. This means that the contexts at the university were determined by the time and location because many periodic activities took place there. TABLE IV lists the correctness of the model generated from a training data set of 600 (150 pieces of data per variable) and the evaluation results when using the training data set and a test data set of 300 (65 pieces of data per variable). As shown in this table, the random forest algorithm achieves the highest levels of accuracy—above approximately 0.900—for all contexts at the university except for "shopping", which was less than 0.900. We assume the reason for this is that "shopping" is an irregular context based on time and the

fact that the store was located on the second floor of the same building as the cafeteria ("meal"). It is possible to improve the accuracy by using altitude values that the GPS sensor can obtain as an explanatory variable.

B. Effect of Filtering

A threshold-based filtering method was used. The GPS sensor in a mobile phone can obtain values of accuracy as an error range. It is necessary to remove the data with large positioning errors because the GPS sensor measures a lot of noise when the sensor is indoors. The noise data were removed to obtain the information discussed in this paper by using the accuracy values of the mobile phone. The purpose in this experiment was to determine a suitable threshold. Figure 7 shows the experimental results.

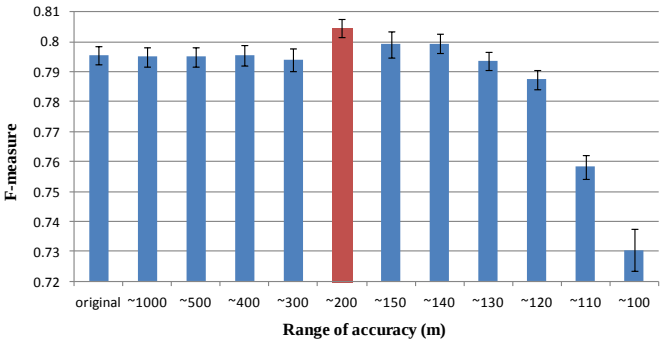


Fig. 7. F-measure when removing data by changing allowable range of accuracy

This experiment used a data set of 400 (research: 100, meeting: 100, meal: 100, shopping: 100) at the university as an original data set. The original data set contained data with a maximum accuracy of 2,373 (m) and a minimum accuracy of 7 (m). The average level of accuracy was 116.92 (m) from all the data. This result was determined on the basis of ten runs using the random forest algorithm. As shown in Figure 7, when removing the data in which the level of accuracy was higher than 200, the context recognition achieved the highest level of accuracy (0.805 overall). At this time, the number of data pieces was 385 (15 pieces were removed). When the number of data pieces was less than 323, the recognition accuracy also began to decrease, as shown in the results for the allowable accuracy range up to 130. The level of accuracy does not improve unless a certain minimum number of data pieces are used while removing any noise. In other words, by setting 200 as the threshold, recognition is possible with a high level of accuracy.

C. Indoor Localization Technology

It is possible to improve the accuracy of daily context recognition by using indoor localization technology. Thus far, other research has proposed indoor localization technologies using Wi-Fi, Japan's Indoor Messaging System (IMES), Dead Reckoning, and others [32]–[36]. Indoor localization is one of the most important elements for daily context recognition. If indoor localization technologies are added to the proposed method, it is possible to recognize more specific contexts for lifelog generations.

For indoor localization in this study, motion sensors were installed in a grid pattern on the ceiling of the authors' laboratory. An experiment was performed to recognize the contexts in the laboratory by using the motion sensor data. Figure 8 shows the experimental results and environment.

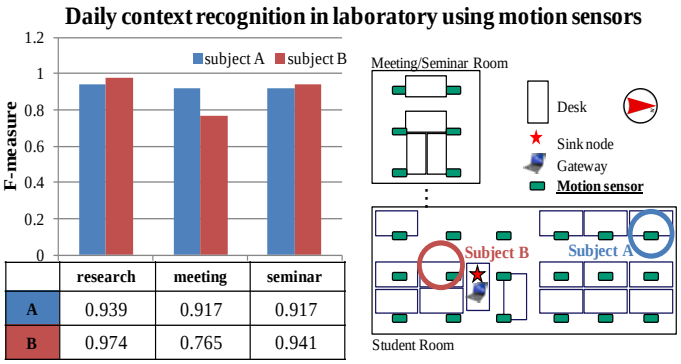


Fig. 8. Results of context recognition in laboratory by using motion sensor data in the experimental environment

Twenty-four motion sensors were installed on the ceiling of the laboratory. Six of them were installed in the meeting/seminar room and the others were installed in the student room. All the motion sensors transmitted sensing data to the gateway by detecting people movement once per minute. The experiment was performed with two participants. Participant A had the following data: 168 "research", 166 "meeting", and 118 "seminar". Participant B had the following data: 341 "research", 104 "meeting", and 20 "seminar". For both participants, the recognition accuracies were above 0.765, and almost all the contexts had a high level of accuracy above 0.900. It is judged that the motion detection from the motion sensors showed regularity. The motion sensor just above the desk of the participants detected many movements when the participants were studying in their seats. The motion sensors installed in the meeting/seminar room detected many movement while the motion sensors installed just above the desk did not detect many movements during a group meeting in the meeting/seminar room. Moreover, when everyone was attending a seminar, the motion sensors installed in the meeting/seminar room detected many movements, and, as expected, the motion sensors installed in the student room hardly detected any movement.

D. Healthcare Service

The management of daily calorie consumption is considered a healthcare service in this paper. Metabolic equivalents (METS) are used to compute the energy consumed during each daily activity. METS is defined as the ratio of a person's working metabolic rate relative to the resting metabolic rate. A person's calorie consumption can be easily calculated by using the METS value as follows:

$$EE(kcal) = 1.05 \times METS \times Duration(hour) \times Weight(kg). \quad (3)$$

Standard tables provide the METS values for a wide range of exercises and activities. TABLE V itemizes some examples of METS values.

TABLE V. METS EXAMPLES

Examples	METS
driving a car, sitting - light office work (research), sitting - meeting, sitting - eating	1.5
eating - talking, walking - less than 2.0 mph	2.0
riding a scooter, stretching	2.5
walking - 2.5 mph, weight lifting	3.0
walking - 5.0 mph, carrying heavy loads	8.0

As shown in TABLE V, METS values can vary. For example, the METS values for walking depend on the speed. Since the proposed method can calculate the speed from the GPS data, it is possible to closely calculate calorie consumption.

Appropriate recommendations are needed to improve a person's lifestyle for the prevention of lifestyle-related diseases. This paper describes some of the following examples of recommendations.

- ✓ If eating a meal is recognized as occurring at an irregular time or not taken three times a day, the user is advised to have a regular eating habit.
- ✓ If riding a scooter is recognized for a short duration many times in a week, the user is advised to walk once in a while.
- ✓ If the expenditure of calories is low for a week, the user is advised to spend more time doing exercises such as stretching and playing sports.
- ✓ If the user is found performing research while seated in a chair for a long time, the user is advised to stand up, walk around and possibly do some light shopping.

It is possible to generate high-quality lifelogs by recognizing the specific contexts in a person's daily life. Several healthcare services can be created by using the lifelogs, enabling people to improve their lifestyles and prevent lifestyle-related diseases.

E. Discussions

With this research, we are interested in context tracking as a simple way to track people's activities, rather than describing and characterizing the taxonomy for a lifelog. Other studies have considered context estimation of daily life by using multiple sensor devices [1], but our originality is our measurement of several variables for daily context recognition by using a commercial device.

The precision of a general-purpose mobile phone for collecting information is often less than that of specialized wearable sensors. However, our estimation technique works well with mobile phone systems, and our advantage is that the mobile phone allows the collecting of daily life data in a way that is simpler than with wearable sensors. This could be an important achievement in the field of consumer electronics.

V. CONCLUSION

This paper described in detail the proposed method of daily context recognition for lifelog generation to prevent lifestyle-related diseases. The proposed method enables the recognition of several user contexts by using machine learning on GPS data from smartphones. We found that the optimal explanatory variables depend on the types of contexts recognized. Most contexts can be recognized by the random forest algorithm with high accuracy. The experimental results demonstrate that it is possible to improve the recognition accuracy by using threshold-based filtering and indoor localization technology. Moreover, lifelogs generated by using the proposed method can help adapt healthcare services in accordance with the user's location and context. In our future work, we will focus on adapting the proposed method for large-scale outdoor facilities by using social data as the explanatory variables. We also intend to consider a low-cost learning technique by taking a non-parametric approach to generalization. For creating context-aware services with consumer devices, we will strive for a well-balanced approach, so that individuals receiving input from these lifelogs do not find the process intrusive.

REFERENCES

[1] M. Ono, K. Nishimura, T. Tanikawa, and M. Hirose, "Neural Network Based Event Estimation on Lifelog from Various Sensors," *IEEE 16th International Conference on Virtual Systems and Multimedia (VSMM)*, pp. 84-87, 2010.

[2] M. Abe, D. Fujioka, and H. Handa, "A Life Log Collecting System Supported by Smartphone to Model Higher-Level Human Behaviors," *2012 Sixth International Conference on Complex, Intelligent, and Software Intensive System (CISIS)*, pp. 665-670, 2012.

[3] D. Feldman, A. Sugaya, C. Sung, and D. Rus, "iDiary: From GPS Signals to a Text-Searchable Diary," *11th ACM Conference on Embedded Networked Sensor System (SenSys)*, Article No. 6, 2013.

[4] World Health Organization, *WORLD HEALTH STATISTICS 2012*, pp. 34-37, 2012.

[5] G. Tanaka, and H. Mineno, "A Method of Estimating Outdoor Situation for Lifelog Generation," *2nd IEEE Global Conference on Consumer Electronics (GCCE2013)*, pp. 361-362, 2013.

[6] O. D. Lara, and M. A. Labrador, "A Mobile Platform for Real-time Human Activity Recognition," *IEEE Consumer Communications and Networking Conference (CCNC)*, pp. 667-671, 2012.

[7] J. B. Gomes, S. Krishnaswamy, M. M. Gaber, P. A. C. Sousa, and E. Menasalvas, "MARS: A Personalised Mobile Activity Recognition System," *IEEE 13th International Conference on Mobile Data Management (MDM)*, pp. 316-319, 2012.

[8] J. R. Kwapisz, G. M. Weiss, and S. A. Moore, "Activity Recognition using Cell Phone Accelerometers," *ACM SIGKDD Explorations Newsletter*, 12(2), pp. 74-82, 2010.

[9] Y. Hattori, S. Inoue, and G. Hirakawa, "A Large Scale Gathering System for Activity Data with Mobile Sensors," *15th IEEE Annual International Symposium on Wearable Computers (ISWC)*, pp. 97-100, 2011.

[10] T. Maekawa, and S. Watanabe, "Unsupervised Activity Recognition with User's Physical Characteristics Data," *15th IEEE Annual International Symposium on Wearable Computers (ISWC)*, pp. 89-96, 2011.

[11] H. Martin, A. M. Bernardos, J. Iglesias, and J. R. Casar, "Activity logging using lightweight classification techniques in mobile devices," *Personal and Ubiquitous Computing*, vol. 17, no. 4, pp. 675-695, 2013.

[12] K. Cho, N. Iketani, H. Setoguchi, and M. Hattori, "Human Activity Recognizer for Mobile Devices with Multiple Sensors," *Symposia and Workshop on Ubiquitous, Autonomic and Trusted Computing (UIC-ATC)*, pp. 114-119, 2009.

[13] R. Liu, T. Chen, and L. Huang, "Research on Human Activity Recognition Based on Active Learning," *9th International Conference on Machine Learning and Cybernetics (ICMLC)*, pp. 11-14, 2010.

[14] N. Ogawa, K. Kaji, and N. Kawaguchi, "Effects of Number of Subjects on Activity Recognition: Findings from HASC2010corpus," *International Workshop on Frontiers in Activity Recognition using Pervasive Sensing (IWFA)*, pp. 48-51, 2011.

[15] N. Kawaguchi et al., "HASC2011corpus: Towards the Common Ground of Human Activity Recognition," *13th International Conference on Ubiquitous Computing (Ubicomp)*, ACM, pp. 571-572, 2011.

[16] N. Kawaguchi et al., "HASC2012corpus: Large Scale Human Activity Corpus and Its Application," *2nd International Workshop of Mobile Sensing: From Smartphones and Wearables to Big Data*, pp. 10-14. 2012.

[17] A. Anjum, and M. U. Ilyas, "Activity Recognition Using Smartphone Sensors," *IEEE Consumer Communications and Networking Conference (CCNC)*, pp. 914-919, 2013.

[18] Y-W. Bai, S-C. Wu, and C-L. Tsai, "Design and implementation of a fall monitor system by using a 3-axis accelerometer in a smartphone," *IEEE Trans. Consumer Electron.*, vol. 58, no. 4, pp. 1269-1275, 2012.

[19] T. Suzuki, and M. Inoue, "Lifestyle Improvement Support System Considering Context of a User," *IEEE International Conference on Consumer Electronics (ICCE)*, pp. 454-455, 2013.

[20] H. Yan, H. Huo, Y. Xu, and M. Gidlund, "Wireless sensor network based e-health system: implementation and experimental results," *IEEE Trans. Consumer Electron.*, vol. 56, no. 4, pp. 2288-2295, 2010.

[21] A. M. Khan, and M. H. Siddiqi, "Promoting a Healthier Life-Style Using Activity-Aware Smartphones," *4th International Conference on Intelligent and Advanced Systems (ICIAS)*, pp. 7-11, 2012.

[22] H. O'Brien, P. van de Ven, J. Nelson, and A. Bourke, "Smartphone Interfaces to Wireless Health Sensors," *12th IEEE International Conference on e-Health Networking Applications and Services (Healthcom)*, pp. 180-186, 2010.

[23] I. Yamada, and G. Lopez, "Wearable Sensing Systems for Healthcare Monitoring," *2012 Symposium on VLSI Technology (VLSIT)*, pp. 5-10, 2012.

[24] R. Albatal, C. Gurrin, J. Zhou, Y. Yang, D. Carthy, and N. Li, "SenseSeer Mobile-Cloud-Based Lifelogging Framework," *IEEE International Symposium on Technology and Society (ISTAS)*, pp. 144-146, 2013.

[25] Z. Li, Z. Wei, W. Jia, and M. Sun, "Daily Life Event Segmentation for Lifestyle Evaluation Based on Multi-Sensor Data Recorded by a Wearable Device," *35th Annual International Conference on the IEEE Engineering in Medicine and Biology Society (EMBC)*, pp. 2858-2861, 2013.

[26] Y.-S. Son, T. Pulkkinen, and J.-H. Park, "Active Monitoring for Lifestyle Disease Patient Using Data Mining of Home Sensors," *IEEE International Conference on Consumer Electronics (ICCE)*, pp. 276-277, 2013.

[27] Z. W. Zhao, L. Liu, Q. Ma, W.-D. Li, and C.-C. Li, "A Machine-to-Machine Based Framework for Diabetes Lifestyle Management," *10th IEEE International Conference on Networking, Sensing and Control (ICNSC)*, pp. 562-567, 2013.

[28] A. M. Khan, and S.-W. Lee, "Need for a Context-Aware Personalized Health Intervention System to Ensure Long-Term Behavior Change to Prevent Obesity," 5*th International Workshop on Software Engineering in Health Care (SEHC)*, pp. 71-74, 2013.

[29] N. Alshurafa et al., "Designing a robust activity recognition framework for health and exergaming using wearable sensors," *IEEE Journal of Biomedical and Health Informatics*, DOI 10.1109/JBHI.2013.2287504, 2013.

[30] S. Suzuki et al., "Development of a topic providing system with inference of behaviors from daily life," *Computer Technology and Application*, vol. 4, no. 3, pp. 144-152, 2013.

[31] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten, "The WEKA data mining software: an update," *ACM SIGKDD Explorations Newsletter*, pp. 10-18, 2009.

[32] H. Liu et al., "Push the Limit of WiFi Based Localization for Smartphones," *18th Annual International Conference on Mobile Computing and Networking (Mobicom'12)*, pp. 305-316, 2012.

[33] K. Wu, J. Xiao, Y. Yi, M. Gao, and L. M. Ni, "FILA: Fine-grained Indoor Localization," *IEEE 2012 INFOCOM*, pp. 2210-2218, 2012.

[34] Y. Sakamoto, H. Arie, T. Ebinuma, K. Fujii, and S. Sugano, "High-Accuracy IMES Localization Using a Movable Receiver Antenna and a Three-axis Attitude Sensor," *International Conference on Indoor Positioning and Indoor Navigation (IPIN)*, pp. 1-6, 2011.

[35] Y. Sakamoto, T. Ebinuma, K. Fuji, and S. Sugano, "GPS-compatible Indoor-positioning Methods for Indoor-outdoor Seamless Robot Navigation," *IEEE International Workshop on Advanced Robotics and its Social Impacts*, pp. 95-100, 2012.

[36] H. Wang et al., "No Need to War-Drive: Unsupervised Indoor Localization," *10th International Conference on Mobile Systems, Applications, and Services (MobiSys'12)*, pp. 197-210, 2012.

Age Estimation Based on AAM and 2*D*-DCT Features of Facial Images

Asuman Günay and Vasif V. Nabiyev
Department of Computer Engineering
Karadeniz Technical University
Trabzon, TURKEY

***Abstract*—This paper proposes a novel age estimation method - Global and Local feAture based Age estiMation (GLAAM) - relying on global and local features of facial images. Global features are obtained with Active Appearance Models (AAM). Local features are extracted with regional 2*D*-DCT (2-dimensional Discrete Cosine Transform) of normalized facial images. GLAAM consists of the following modules: face normalization, global feature extraction with AAM, local feature extraction with 2*D*-DCT, dimensionality reduction by means of Principal Component Analysis (PCA) and age estimation with multiple linear regression. Experiments have shown that GLAAM outperforms many methods previously applied to the FG-NET database.**

***Keywords*—2D-DCT; AAM; Age estimation; PCA; Regression**

I. INTRODUCTION

The wide-ranging topic of facial image (FI) processing has been receiving considerable interest lately because of its real world applications such as forensic art, electronic consumer relationship management, security control and surveillance, cosmetology, entertainment and biometrics. In the FI context, age recognition (or estimation) has been demanding growing attention. Age synthesis, also called age progression is defined as re-rendering FIs with natural and rejuvenating effects. Age estimation (AE) can be defined as the process of associating a FI automatically with an exact age or age group.

In order to facilitate AE, suitable facial representations are necessary. Otherwise, even the most robust classifiers will fail due to the inadequacy of the domain where the feature recognition is done [1]. Hence, the design of face recognition systems requires careful selection of the face feature recognition (FFR) domain. Some issues that should be contemplated are: (i) good discrimination of different people with tolerance to discrepancies inside a class; (ii) FFR must be effortlessly performed from raw face images to speedup processing; and (iii) the FFR must lie in a low dimensional space, in order to facilitate the implementation of the classifiers.

The FI characteristics make the FFR problem very difficult to solve. The most important hindrances are: (1) AE is not a standard classification problem; (2) a large aging database, especially a chronometrical image series of an individual is often hard to collect; and (3) real world age progression displayed on faces is uncontrollable and personalized.

Several techniques have been suggested to represent FIs for recognition purposes, but there is still no consensus on the best when it comes to age recognition/classification. Appearance-based techniques consider an FI as a 2*D* array of pixels and focus on deriving descriptors for face appearance without precise geometrical representations. Holistic (nonparametric) methods such as the Principal Component Analysis (PCA) [2] and the Linear Discriminant Analysis (LDA) [2, 3] along with more recent approaches like 2*D*-PCA [3, 4] and 2*D*-LDA [3] have been broadly studied. Other important approaches handle local descriptors, as for example, Scale Invariant Feature Transform (SIFT), and Affine-SIFT (ASIFT) [5, 6], and they have gained increasing awareness thanks to their robustness to problems akin to pose and illumination alterations [7, 8].

In this paper, we propose a novel Global and Local feAture based Age estiMation (GLAAM) method as shown in Fig. 1. The input images are normalized and the local features are extracted using regional 2*D*-DCT (2-dimensional Discrete Cosine Transform). Global features are obtained with Active Appearance Models (AAM). After feature extraction, dimensionality reduction is performed with PCA. Then, AE is cast as a regression problem. Our method uses global and local considerations and does not rely on a complex Bayesian framework [9]; besides that, it is simple and relatively fast when compared to other ones.

A survey on AE is given in the next section. Section 3 introduces the proposed method for AE including preprocessing, feature extraction, dimensionality reduction and regression modules. In Section 4, experimental results are given and Section 5 concludes the paper.

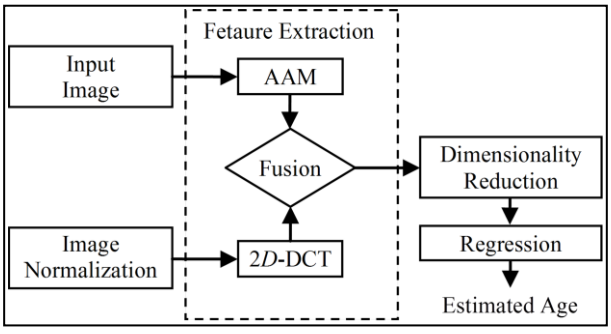


Fig. 1. System structure

II. AGE ESTIMATION METHODS

AE can be seen either as a multi-class classification problem or a regression problem. The existing AE systems typically consist of an age image representation and an AE module. Age image representation techniques rely often on shape-based and texture-based features that were extracted from FIs. They can be grouped under the topics of Anthropometric Models, AAM, AGing pattErn Subspace (AGES), Age Manifold and Appearance Models. Then, AE can be performed with age group classification or regression methods. In recent studies hybrid systems using classification and regression techniques together are presented [10]. Robust multi-instance regressor learning algorithm is also used to build a universal human age estimator, based on facial information [11].

The anthropometric model based representations only consider the facial geometry. The earliest paper published in the area of age classification relying on facial geometry was the work by Kwon and Lobo [12]. They used craniofacial development theory [13] which uses a mathematical model to describe the growth of a person's head. They computed six ratios of distances on frontal images to separate babies from adults. This AE method can only deal with young ages since the human head shape doesn't change too much in its adulthood. So Kwon and Lobo [12] used wrinkle information to separate young adults from senior adults. They used a very small database containing 45 images in their experiments. Later on Horng et al. [14] and Dehshibi and Bastanfard [15] proposed age classification methods using distance ratios based on face anthropometry as geometric features and wrinkle information as texture features.

AAM [16, 17] based approaches consider both shape and texture rather than just the facial geometry as in the anthropometric model based methods. AAM uses a statistical model of object shape and appearance to synthesize a new image throughout a training stage which provides to the training supervisor a set of images and coordinates of landmarks existing in all of the images. AAMs represent a familiar group of algorithms for fitting shape models to images. Training a model requires labeling a database of images where a set of locations called landmarks typify the object group in question. The formulation in [17] chooses a linear and generative model, i.e. an explicit model of the input data has to be provided. This leads to an iterative Gauss-Newton type procedure, where the error between the current image features and those synthesized using the current location of the model in the image are used to derive additive updates to the shape model parameters. Nonetheless, the computational load is heavy, since an explicit image feature model must be stated and evaluated at each algorithm iteration [16]. Lanitis et al. [18] extended AAMs for aging faces by proposing an aging function, *age*=*f*(*b*) which explains the variation in age. But they have to deal with each aging face image separately. Kohli et al. [19] extracted feature vectors from images using AAMs and used ensemble of classifiers trained on different dissimilarities to distinguish between child/teen-hood and adulthood. By using the different aging functions, accurate age of the classified image is estimated. Chao et al. [20] proposed an age estimation method using AAM features. Their approach is based on label sensitive learning and age-oriented regression.

Geng et.al. [21, 22] proposed a method called AGES that defines a sequence of personal face images of the same person sorted in temporal order. Then, a specific aging pattern is learned for each individual. AGES method can synthesize the missing age images by using an expectation maximization-like iterative learning algorithm.

Instead of learning a specific aging pattern for each individual as in AGES, age manifold [23] methods can learn a common aging trend or pattern from many individuals at different ages. This kind of aging pattern learning helps face aging representation. Age manifold utilizes a manifold embedding technique to discover the aging trend in a low dimensional domain from many face images at each age. Thus, the mapping from the image space to the manifold space can be done either by linear or by nonlinear functions [24-28], such as Y=P(X, L), where X is the image space sampled by a set of face images, $X = [x_i : x_i \in R^D]_{i=1}^n$. A ground truth set $L = [l_i : l_i \in N]_{i=1}^n$ associated with images provides the age labeling. $Y = [y_i : y_i \in R^d]_{i=1}^n$ with $d \leq D$ is the low-dimensional representation of X in the embedded subspace. Compared with AGES, all ages of different individuals could be used together in age manifold. The only requirement for the age manifold representation is that the size of the training data set should be large enough in order to learn the embedded manifold with statistical sufficiency.

Appearance models are mainly focused on aging-related facial feature extraction. Both global and local features were used in existing AE systems. Fukai et al. [29] applied Fast Fourier Transform to extract feature spectrum from facial appearance and used genetic algorithm for feature selection. Local Binary Patterns (LBP) have been used as effective texture descriptors for appearance feature extraction [30]. Ju and Wang [31] selected regions with Adaboost that vary with the aging process and used LBP histograms of these regions for AE. Gabor features have also been tried on AE tasks and proved to be more effective than LBP [32]. Guo et al. introduced the biologically inspired features (BIF) to model the aging process on faces [33]. The first layer is created using Gabor filters on facial images. In the second layer, they have proposed to use a novel operator. Their experimental results have shown significant improvement in age estimation accuracy over previous methods. El-Dib and Onsi [34] used BIF to analyze the different facial parts: eye wrinkles, internal face and whole face. According to their analysis, the eye wrinkles contain the most important aging features compared with others. Li et al. [35] and Han et al. [36] also used BIF in their age estimation frameworks.

III. THE GLOBAL AND LOCAL FEATURE BASED AGE ESTIMATION

This paper introduces an innovative AE method – known as GLAAM – relying on local and global facial features of images. Local features are extracted using regional 2*D*-DCT of normalized FIs and the global features are produced by AAMs. This method consists of the following modules: (1) face normalization, (2) global feature extraction with AAM

and local feature extraction with 2*D*-DCT, (3) dimensionality reduction with PCA, and (4) AE with multiple linear regression.

A. Face Normalization

Since shape and local variations of images during aging suffer an evident influence from rotation, scaling and translation, all the images have to be compatible with a common shape model produced by means of a training set of samples. In order to train the shape model, each image is represented by the coordinates of 68 landmark points (Fig. 2-a). Then, the statistical shape model is trained and all images are warped to the mean shape, so that shape variations within the training set are eliminated. The warping process employs affine transformation and Delaunay triangulation (Fig. 2-b,c). Since the FIs vary in head pose, the warped images are inclined to the left as shown in Fig. 2-c. So we rotate these images and cropped the main face part and scaled to the size of 88×88 (Fig. 2-d). Thus we almost eliminate the unreasonable regions for feature extraction. In the local feature extraction phase the images are divided into blocks each having 8×8 block size. So we scale images to the size of 88×88 which is large enough for age related feature extraction. This image size is also efficient in terms of computational costs.

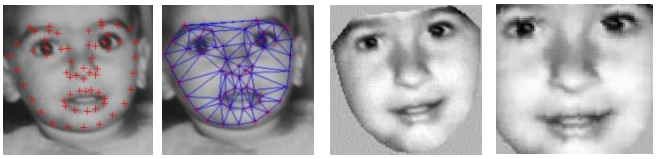
B. Feature Extraction

The feature extraction module consists of two phases: global feature extraction with AAM and local feature extraction with 2*D*-DCT computation. These steps will be explained in detail in the following sections.

1) Global Feature Extraction with AAM: AAM is a statistical shape and appearance model of FIs [17]. These models are generated by combining a model of shape variations with a model of the appearance variations in a shape-normalized frame. A statistical shape model can be generated with a training set of face images labeled with landmark points as shown in Fig. 2-a. Let us represent all the landmark points of training images by $X = [x_i : x_i \in R^D]_{i=1}^n$. The mean shape is produced with taking the mean of the landmark points in the training set as $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$. Then, PCA is applied to the data to extract the main principal components along which the training set varies from the mean shape. If the total scatter matrix S is defined as $S = \sum_{i=1}^{n}(x_i - \bar{x})(x_i - \bar{x})^T$, the projection is chosen to maximize the determinant of the total scatter matrix of the projected samples as $arg\ max_{P_s}|P_s{}^T S P_s|$. P_s is the set of eigenvectors of S corresponding to the d largest eigenvalues. Then a linear transformation maps the D-dimensional data space into a d-dimensional parameter space where $d \leq D$. The shape parameters are defined by linear formulation as $b_s = P_s{}^T X$. As a result, any training set of images can be approximated by,

$$x = \bar{x} + P_s b_s \quad (1)$$

where $\bar{x}$ is the mean shape, P_s is a set of orthogonal principal modes of variation and b_s is a set of shape parameters.



(a) (b) (c) (d)

Fig. 2. a) Example of face image labeled with 68 landmark points b) Result of the Delaunay triangulation used in warping process c) Result of face normalization d) Facial image used in the local feature extraction phase

To build a statistical appearance model, each image has to be normalized, so that its control points match the mean shape using a Delaunay triangulation (Fig. 2.b, c). Then, the gray-level intensities within a pre-specified image region are stacked to form vector g are used for training an intensity model. By applying PCA to the gray level intensities, a linear model is obtained as follows:

$$g = \bar{g} + P_g b_g \quad (2)$$

where $\bar{g}$ is the mean gray-level vector, P_g is a set of orthogonal modes of variation and b_g is a set of gray-level parameters. The shape and appearance of any image can be summarized by the vectors b_s and b_g. Since there may be correlations between the shape and gray-level variations, a further PCA is applied to them and, finally, the combined shape and appearance parameters are obtained. For the intensity model, approximately 7000 gray-level intensities in the facial region of the corresponding shape-free image are used to represent the training samples. The resulting combined shape and intensity model requires 277 model parameters to explain 95 percent of the variance in the training set. These model parameters are used as a global descriptor of FI's.

2) Local Feature Extracting with 2D-DCT: DCT is an invertible linear transform that can express a finite sequence of data points in terms of a sum of cosine functions. The original signal is converted to the frequency domain by applying the direct DCT transform and it is possible to convert back the transformed signal to the original domain by applying the inverse DCT transform (IDCT). After the original signal has been transformed, its DCT coefficients reflect the importance of the frequencies that are present in it.

The 2*D*-DCT is commonly used as a pre-processing step in face recognition, because it attenuates the problems created by changes due to illumination angles, face occlusions, colors and pose [37]. Using the face images directly for recognition purposes resulted in inefficiencies because of the high information redundancy and correlation in such images. Therefore DCT is widely used as a feature extraction and compression method in various applications due to its properties such as de-correlation, energy compaction, separability and orthogonality [38]. All these properties lead us to use 2*D*-DCT in AE field.

De-correlation: The principle advantage of image transformation is the removal of redundancy between neighboring pixels. This leads to un-correlated transform coefficients which can be encoded independently without compromising coding efficiency.

Energy-compaction: Efficacy of a transformation scheme can be directly gauged by its ability to pack input data into as few coefficients as possible. This allows the quantizer to discard coefficients with relatively small amplitudes without introducing visual distortion in the reconstructed image. DCT exhibits large variance distribution in a small number of coefficients for highly correlated images such as face images. In other words DCT packs energy in the low frequency regions. Therefore some of the high frequency content can be discarded without significant quality degradation.

Separability: The 1D-DCT (1 dimensional DCT) transform can be represented as:

$$F(k) = \alpha(k) \sum_{n=0}^{N-1} f(n) \cos[\frac{(2n+1)k\pi}{2N}] \quad (3)$$

The 2*D*-DCT transform can be expressed as:

$$F(k_1, k_2) = 1D - DCT_{y-dir}(1D - DCT_{x-dir}) \quad (4)$$

This property has the principle advantage that $F(k_1, k_2)$ can be computed in two steps by successive 1D operations on rows and columns of an image.

Orthogonality: In pattern recognition techniques to make the model computationally efficient, transform orthogonality is as important as the class separation in applications like face recognition. Unlike Gabor elementary functions, which are a set of overlapping functions and not mutually orthogonal, the DCT basis functions are orthogonal. In addition to its de-correlation characteristics, this property renders some reduction in the pre-computation complexity.

The 2*D*-DCT definition is given in (5).

$$X_{k_1,k_2} = \sum_{n_1=0}^{N_1-1} \sum_{n_2=0}^{N_2-1} x_{n_1,n_2} \cos\left[\frac{\pi}{N_1}\left(n_1 + \frac{1}{2}\right)k_1\right] \cos\left[\frac{\pi}{N_2}\left(n_2 + \frac{1}{2}\right)k_2\right] \quad (5)$$

In the proposed method, the normalized 88×88 images are divided into 11×11 blocks each having dimension of 8×8 and 2*D*-DCT is applied to them. This block size was adopted by the JPEG compression standard [39]. In the developmental phases the processing of larger blocks was seen as being prohibitively slow for the computer to execute. Also the experts observed that the use of larger blocks did not result in appreciably greater compression and quantization artifacts become more visible as the block size increases.

In practice for a wide range of images and viewing conditions, 8×8 has been found to be the optimum DCT block size and is specified in most current coding standards. After applying 2*D*-DCT we have 64 coefficients for each block. To eliminate the high frequency coefficients, quantization is performed. Coefficients are arranged in a vector following a zigzag fashion and the first 21 coefficients to represent that image block. Hence, the dimension of a local feature vector is 11×11×21=2541.

After the global and local features are extracted, they are combined in a single vector in order to perform dimensionality reduction with PCA. To combine the global and local features, a feature level fusion approach is used. For this purpose, the feature vectors are normalized by the *z*-score normalization as:

$$\hat{f}_{i,j} = \frac{f_{i,j} - \mu_j}{\sigma_j} \quad \text{with} \quad j=1,2 \text{ and } i=1,\ldots,n \quad (6)$$

where n is the number of images, $f_{i,j}$ is the j-th feature vector of i-th image and μ_j, $\sigma_j$are the mean and standard deviation of feature vector $f_{i,j}$, respectively. Then, the fused feature vector is created by concatenating the normalized global and local feature vectors as follows:

$$f_{i,fused} = [\hat{f}_{i,1}\ \hat{f}_{i,2}] \quad i=1,\ldots,n \quad (7)$$

C. Dimensionality Reduction

After the feature extraction module, PCA is performed in order to find a lower dimensional subspace which carries significant information for AE. Then, high-dimensional feature vectors are projected onto a low-dimensional subspace in order to improve the efficiency. Using this technique the p-dimensional feature vector f is transformed into a d-dimensional vector y with $d \leq p$.

The PCA method finds the embedding that maximizes the projected variance given below.

$$W_{opt} = \arg\max_{\|W\|=1} W^T S W \quad (8)$$

In (8) $S = \sum_{i=1}^{n} (f_i - \bar{f})(f_i - \bar{f})^T$ is the scatter matrix, and $\bar{f}$ is the mean vector of $\{f_i\}_{i=1}^{n}$. The solution of this problem is given by the set of d eigenvectors associated to the d largest eigenvalues of the scatter matrix. Once the projection subspace is determined, training and testing images were projected on it. The low dimensional representation of feature vectors is calculated with $y_i = W^T f_i$ allowing thus dimensionality reduction.

D. Regression

After finding a lower dimensional representation of facial images, we recast the AE problem as a multiple linear regression as follows:

$$age = F(M) : \leftrightarrow \hat{L} = \hat{F}(Y) \quad (9)$$

where $\hat{L}$ denotes the estimated age label, $F(\cdot)$ the unknown regression function, and $\hat{F}(.)$ is the estimated regression function. The corresponding matrix formulation is

$$L = \tilde{Y}B + e,\ Var(e) = \sigma^2 I \quad (10)$$

where L is the age label vector. $\tilde{Y}$ is a known matrix including a column of 1s for the intercept and observed values. The vector B is the unknown parameter vector which we need to estimate during the learning stage. The error vector e consists of unobservable random variables, and assumed to have zero mean and uncorrelated with common variance σ^2. In fitting model, B is estimated by ordinary least squares $\hat{B} = (\tilde{Y}'\tilde{Y})^{-1}\tilde{Y}'L = HL$ or robust regression, and the fitted value of L is given by,

$$\hat{L} = \tilde{Y}\hat{B}. \quad (11)$$

The vector of residuals is $\hat{e} = L - \hat{L}$, with $E(\hat{e}) = 0$ and $Var(\hat{e}) = \sigma^2(I - H)$. The age regression function used in this study is a linear function given by

$$\hat{l} = \hat{\beta}_0 + \hat{\beta}_1{}^T y \quad (12)$$

where $\hat{l}$ is the estimate of age, $\hat{\beta}_0$ is the offset, $\hat{\beta}_1$ is the weight vector and *y* is the extracted feature vector.

IV. Experiments and Results

In this paper, the FG-NET Aging Database [40] is used to train and test the proposed method. This database contains 1,002 face images from 82 subjects with approximately 10 images per subject. The ages in the database are distributed in a wide range from 0 to 69. The age distribution of the FG-NET database is given in Table 1. One can see from the table that the images are not distributed uniformly.

A typical aging sequence from the FG-NET database is shown in Fig. 3-a. Besides the aging variation, most aging sequences display variations in pose, illumination, facial expression, occlusion, etc. Although these variations may increase computational complexity, all the images have been used in the experiments to avoid restrictions.

The normalization phase determines the mean shape from the 68 landmarks of training samples. Next, all images are warped to the mean shape (Fig. 3.b) using affine transformation and Delaunay triangulation and scaled to the size of 88x88. Furthermore, each image is represented with 277 AAM model parameters that are used as global face features.

In the local feature extraction step, the normalized 88x88 images are divided into 11x11 blocks each having 8x8 size and 2*D*-DCT is applied to them. After 2*D*-DCT computing, we have 64 coefficients for each block. To eliminate the high frequency coefficients, quantization is performed. Coefficients are arranged in a vector according to a zigzag fashion. In this phase the determination of the number of DCT coefficients is done experimentally. For this purpose, the AE performance of different number of coefficients is calculated and the results are listed in Table 2. We can see from Table 2 that using the 21 DCT coefficients gives better results than other ones. So the first 21 coefficients are selected to represent each image block. Hence, the dimension of the local feature vector is 11x11x21 (121 blocks with 21 entries/block).Then global and local feature vectors are normalized according to their mean and standard deviation and concatenated into a single vector. After that, a low dimensional age manifold is learned with PCA. AE is performed with multiple linear regression in the low dimensional space.

TABLE I. Age Range Distribution of the Images in FG-NET Database

Age Range	FG-NET (%)
0-9	37.03
10-19	33.83
20-29	14.37
30-39	7.88
40-49	4.59
50-59	1.50
60-69	0.80

(a)

(b)

Fig. 3. (a) Typical aging face sequence in FG-NET Aging Database; and (b) Normalized face sequence

TABLE II. The Estimation Results of Different Number of DCT Coefficients

Number of DCT Coefficients	MAE
1	7,25
3	6,85
6	6,65
10	6,43
15	6,26
21	6,18
28	6,20
36	6,19

Performance evaluation is done by means of a cross-validation variant known as the leave-one-person-out (LOPO), i.e., in each fold the images of one person are used as test set and those of the others are used as the training set. As FG-NET contains face images from 82 subjects, after 82 folds, each subject has been used as test set once, and the final results are calculated based on all estimations. This scenery is very close to real life applications and, hence, it is very adequate for testing. The Mean Absolute Error (MAE) has been chosen as a metric for performance comparison. MAE is defined as the average of the absolute error between the recognized labels and the ground truth labels as follows:

$$MAE = \sum_{i=1}^{N} \left|\hat{l}_i - l_i\right| / N_t \quad (13)$$

where $\hat{l}_i$ is the recognized age for the *i*th testing sample, l_i is the corresponding ground truth, and N_t is the total number of the testing samples.

The estimation results of earlier methods and GLAAM are listed in Table 3. As one can infer from Table 3, GLAAM achieves better results than earlier methods like WAS, AAS, KNN and AGES on the FG-NET database. DCT encodes facial texture and edge information in the frequency domain. Moreover local appearance information is captured using the block based DCT, but the global ones are ignored. So we use AAM, because it encodes the geometrical and global facial texture information in spatial domain. These feature sets capture differential complementary information. The combination of these feature vectors outperforms the AE accuracy of each one of the feature vector alone.

We also investigate the AE performance of our method in various age ranges. The estimation results are given in Table 4. From Table 4 we can observe that, GLAAM outperforms the AE accuracy of global features and local features alone, almost in all age ranges. As the age variation in the age range increased, the effectiveness of our method became more outstanding as shown in Fig. 4.

V. CONCLUSION

In this paper, an AE method relying on an AAM model named GLAAM has been introduced. Its main contribution is a set of parameters accounting for both global texture features as well as local features of FI's. Locality is preserved by regional DCT coefficients and this is the main advantage/contribution of GLAAM over its competitors because DCT captures more accurately local features in FIs. Moreover, the proposed method is simple and relatively fast when compared to other ones used as benchmark, because 2*D*-DCT is recast and computed by means of 1*D*-DCT operations.

Shape variations are eliminated via normalization with respect to mean shape obtained from a training set consisting of FIs. Furthermore, these local features are combined with global features of images. Experimental results using the FG-NET aging database show that GLAAM is better than earlier methods. However, there is plenty of room for research, since there are methods that do not require normalization such as SIFT and ASIFT. An extra improvement in GLAAM would be the use of Principal Component Regression, since it combines PCA and Regression in the same stage.

TABLE III. THE COMPARISON OF ESTIMATION RESULTS ON FG-NET DATABASE

Methods	MAEs
WAS [17]	8.06
AAS [17]	14.83
KNN [18]	8.24
BP [18]	11.85
SVM [18]	7.25
AGES [17]	6.77
$AGES_{lda}$ [18]	6.22
AAM	6.07
2*D*-DCT	6.18
AAM+2*D*-DCT	5.39

TABLE IV. AGE ESTIMATION RESULTS AT DIFFERENT AGE RANGES IN FG-NET DATABASE

Age Group	#img.	AAM	2*D*-DCT	AAM+2*D*-DCT
0-9	371	2,63	1,35	1,62
0-19	710	2,81	2,46	2,34
0-29	854	3,77	3,45	3,24
0-39	933	4,77	4,59	4,20
0-49	979	5,51	5,48	4,91
0-59	994	5,89	5,85	5,23
0-69	1002	6,07	6,18	5,39

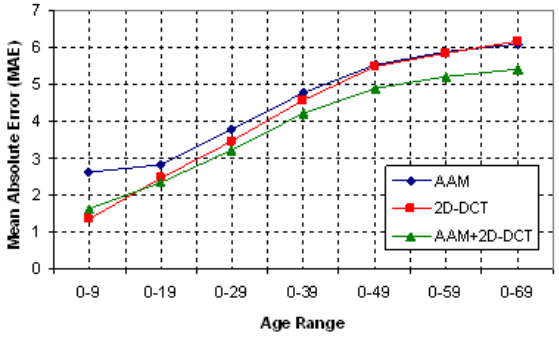


Fig. 4. MAEs at different age ranges in FG-NET

REFERENCES

[1] A. Hadid, M. Pietikäinen and T. Ahonen, "A discriminative feature space for detecting and recognizing faces", Proc. of Computer Vision and Pattern Recognition, pp.797-804, 27 June-2 July 2004.

[2] W. Chen, M. J. Er and S. Wu, "PCA and LDA in DCT domain", Pattern Recognition Letters, vol. 26, no. 15, pp. 2474-2482, 2005.

[3] M. Li and B. Yuan, "2*D*-LDA: A statistical linear discriminant analysis for image matrix", Pattern Recognition Letters, vol. 26, no. 5, pp. 527-532, 2005.

[4] H. Yu and M. Bennamoun, "1D-PCA, 2*D*-PCA to nD-PCA", 18th International Conference on Pattern Recognition (ICPR'06), pp. 181-184, 2006.

[5] D. G. Lowe, "Distinctive image features from scale-invariant keypoints", International Journal of Computer Vision, vol. 60, no. 2, pp. 91–110, 2004.

[6] J.-M. Morel and G. Yu, "ASIFT: A new framework for fully affine invariant image comparison", SIAM Journal on Imaging Sciences, vol. 2, no. 2, pp. 438-469, 2009.

[7] M. Hamouz, J. Kittler, J.-K. Kamarainen, P. Paalanen, H. Kalviainen and J. Matas, "Feature-based affine-invariant localization of faces", IEEE Transactions on Pattern Recognition and Machine Intelligence, vol. 27, no. 9, pp. 1490-1495, 2005.

[8] M. Bicego, A. Lagorio, E. Grosso and M. Tistarelli, "On the use of SIFT features for face authentication", Proc. of the Computer Vision and Pattern Recognition Workshop (CVPRW'06), pp. 17-22, June 2006.

[9] P. Paalanen, J-K. Kamarainen, J. Ilonen and H. Kälviäinen, "Feature representation and discrimination based on Gaussian mixture model probability densities-practices and algorithms", Pattern Recognition, vol. 39, no. 7, pp. 1346-1358, 2006.

[10] S. E. Choi, Y. J. Lee, S. J. Lee and K. R. Park, "Age estimation using a hierarchical classifier based on global and local facial features", Pattern Recognition, vol. 44, no. 6, pp. 1262-1281, June 2011.

[11] B. Ni, Z. Song and S. Yan, "Web image and video mining towards universal and robust age estimator", IEEE Transactions on Multimedia, vol. 13, no. 6, pp. 1217-1229, December 2011.

[12] Y. H. Kwon and N. V. Lobo, "Age classification from facial images", Computer Vision and Image Understanding, vol. 74, no. 1, pp. 1-21, April 1999.

[13] T. R. Alley, Social and Applied Aspects of Perceiving Faces, Lawrence Erlbaum Associates, Hillsdale, NJ, 1988.

[14] W.-B. Horng, C.-P. Lee and C.-W. Chen, "Classification of Age Groups Based on Facial Features", Tamkang Journal of Science and Engineering vol. 4, no.3, pp. 183-192, 2001.

[15] M. M. Dehshibi and A. Bastanfard, "A new algorithm for age recognition from facial images", Signal Processing, vol. 90, no.8, pp. 2431-2444, 2010.

[16] S. Baker and I. Matthews, "Lucas-Kanade 20 years on: A unifying framework", International Journal of Computer Vision, vol. 56 , no. 3, pp. 221-255, 2004.

[17] T. Cootes, G. Edwards and C. Taylor, "Active appearance models", IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 23, no. 6, pp. 681-685, Jun 2001.

[18] A. Lanitis, C. Taylor and T. Cootes, "Toward automatic simulation of aging effects on face images", IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 24, no. 4, pp. 442-455, April 2002.

[19] S. Kohli, S. Prakash and P. Gupta, "Hierarchical age estimation with dissimilarity-based classification", Neurocomputing, vol. 120 pp. 164-176, 2013.

[20] W. -L. Chao, J. -Z. Liu and J. -J. Ding, "Facial age estimation based on label-sebsitive learning and age oriented regression", Pattern Recognition, vol. 43, pp. 628-641, 2013.

[21] X. Geng, Z. H. Zhou, Y. Zhang, G. Li and H. Dai, "Learning from facial aging patterns for automatic age estimation", Proc. of ACM Conference on Multimedia, pp. 307-316, 2006.

[22] X. Geng, Z. H. Zhou and K. S. Miles, "Automatic age estimation based on facial aging patterns", IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 29, no. 12, pp. 2234-2240, December 2007.

[23] Y. Fu, Y. Xu and T. S. Huang, "Estimating human age by manifold analysis of face pictures and regression on aging feature"s, Proc. IEEE International Conference on Multimedia and Expo, pp. 1383-1386, 2-5 July 2007.

[24] D. Cai, X. He, J. Han and H.-J. Zhang, "Orthogonal laplacianfaces for face recognition", IEEE Transactions on Image Processing, vol. 15, no. 11, pp. 3608-3614, 2006.

[25] Y. Fu and T. S. Huang, "Human age estimation with regression on discriminative aging manifold", IEEE Transactions on Multimedia, vol. 10, no. 4, pp. 578-584, June 2008.

[26] G. Guo, Y. Fu, T. S. Huang and C. R. Dyer, "Locally adjusted robust regression for human age estimation", IEEE Workshop on Applications of Computer Vision (WACV'08), pp. 1-6, 7-9 Jan 2008.

[27] G. Guo, Y. Fu, C. R. Dyer and T. S. Huang, "Image-based human age estimation by manifold learning and locally adjusted robust regression", IEEE Transactions on Image Processing, vol. 17, no. 7, pp. 1178-1188, July 2008.

[28] J. Lu and Y. -P. Tan, "Ordinary Preserving Manifold Anaysis for Human Age and Head Pose Estimation", IEEE Transactions on Human-Machine Systems, vol.43, no.2, pp. 249-258, 2013.

[29] H. Fukai, H. Takimoto, Y. Mitsukura and M. Fukumi, "Apparent age estimation system based on age perception", Proc. SICE 2007 Annual Conference, pp. 2808-2812, , 17-20 Sept 2007.

[30] T. Ahonen, A. Hadid and M. Pietikainen, "Face description with local binary patterns: Application to face recognition", IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 28, no.12, pp. 2037-2041, December 2006.

[31] C. H. Ju and Y. H. Wang, "Automatic age estimation based on local feature of face image and regression", 2009 International Conference on Machine Learning and Cybernetics, pp. 885-888, 12-15 July 2009.

[32] F. Gao and H. Ai, "Face age classification on consumer images with gabor feature and fuzzy LDA method", Proc. of 3rd International Conference on Advances in Biometrics (LNCS'5558), pp. 132-141, 2-5 June 2009.

[33] G. Guo, G. Mu, Y. Fu and T. S. Huang, "Human Age Estimation Using Bio-Inspired Features", Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, pp. 112-119, 2009.

[34] M. Y. El Dib and H. M. Onsi, "Human age estimation framework using different facial parts", Egyptian Informatics Journal, vol.12, no.1, pp. 53-59, 2011.

[35] H. Han, C. Otto, X. Liu and A. Jain, K., "Demographic Estimation from Face Images: Human vs. Machine Performance", IEEE Transactions on Pattern Analysis and Machine Intelligence, 2014.

[36] C. Li, Q. Liu, W. Dong, X. Zhu, J. Liu and H. Lu, "Human Age Estimation Based on Locality and Ordinal Information", IEEE Transactions on Cybernetics, 2014.

[37] W. Chen, M. J. Er and S. Wu, "Illumination compensation and normalization for robust face recognition using discrete cosine transform in logarithm domain", IEEE Transactions on Systems, Man and Cybernetics, Part B: Cybernetics, vol. 36, pp. 458-466, 2006.

[38] K. Rao and P. Yip, Discrete cosine transform, Algorithms, Advantages, Applications, Academic Press, Boston, MA, 1990.

[39] W. Pennebaker and J. Mitchell, JPEG: Still image data compression standard, Kluwer Academic Publishers, Norwell, MA, 1992.

[40] FG-NET aging database. [Online]. Available: http://sting.cycollege.ac.cy/ alanitis/fgnetaging/

Personal Health Book Application for Developing Countries

Seddiq Alabbasi, Andrew Rebeiro-Hargrave, Kunihiko Kaneko, Ashir Ahmed and Akira Fukuda
Department of Advanced Information Technology Kyushu University
Fukuoka, Japan

***Abstract*—We introduce a Personal Health Book application that is used as a portable repository for Personal Health Records (PHR) in order to alleviate healthcare organizational problems in developing countries. The Personal Health Book application allows low literate people to access and carry their own medical history from a rural healthcare provider to an urban healthcare provider. This will improve the efficiency of medical care and lower costs for health clinics in underserved areas. This paper introduces a software application that can be ported onto a USB Smart Card or/and managed by smartphone or personal computer connected to cloud computing environment. The Portable Health Book application aims to ease the problem of interoperability between health clinics by accepting any file format and contents and applies a decomposed database to categorize, group and reorganize the data. Querying the application's database, the consumer can create a unified report presentation that is understandable by the consumer, family, and healthcare provider. We tested the Personal Health Book framework by importing PHRs in an extensible markup language (XML) format with a basic structure, without checking the PHR content from the Grameen Portable Health Clinic database in Bangladesh and from different departments from a hospital in Japan. The Personal Health Book was able to generate a human readable output as its database reorganize and store any type of PHR including sensor device data.**

Keywords*—Personal health records; Patient centered healthcare; Database design; Developing countries; Extensible markup language

I. INTRODUCTION

The introduction of an affordable and small Personal Health Book software application that stores Personal Health Records can help alleviate healthcare organizations' problems in developing countries.

Personal Health Records (PHR) are a set of computer-based tools that allow people to access and coordinate their lifelong health information and make appropriate parts of it available to those who need it [1,2]. PHR systems were developed in the late 1990s to target patients who were travelling and needed healthcare and for situations where patients were not able to provide their health information [3,4], or have communication difficulties.

Currently, three types of PHRs have become available [5]. These include: stand-alone formats (PC, USB drive), where consumers store health information on personal computers but lack the ability to exchange information between consumers and healthcare provides; Web formats that are managed by third parties (such as Microsoft Healthvalut, Dossia Consortium) and allow consumers to maintain their health information on private online accounts accessed by a login ID and password; Integrated PHRs with Electronic Health Records (EHR) where a healthcare provider (such as MyHealthVet of the US Department of Veteran Affairs) combines patient entered content with EHR data [6].

Despite significant interest and anticipated benefits for consumers and healthcare providers, the overall adoption of PHRs remains relatively low [7]. A study on MyHealthVet PHR system showed the authentication and secure messaging had important consequences for access, communication, patient self-report and patient/provider relationship. From a literature review of 28 articles the common challenges for the use of PHRs include: data accuracy, data privacy and security, digital divide, and literacy issues.

Accepting the PHR challenges, the concept of a Personal Health Book (PHB) has been introduced to the literature. Similar to Web formats and online PHR health vaults, the Personal Health Book-based healthcare model is a cloud-based service where consumer PHRs are stored at a remote server and shared with healthcare providers (such as a pharmacy) whom the consumer authorizes and who has the capacity to import data from other information sources [8,9]. An alternative approach is to revisit the "stand-alone approach" and let the consumer have a small "personal health book" software application that stores different types of PHRs from different healthcare providers. This is similar to a diary that houses "personal lifetime data" and is an application with standard interfaces and its own database [10].

A simple digital Personal Health Book that stores PHRs can overcome the digital divide and improve the efficiency of healthcare organization in developing countries by enabling people to carry their own medical history. Currently, millions of low-income people frequently travel from rural communities to urban centres and do not have any access to their own medical records. When patients revisit health clinics, doctors often waste their time and resources finding and organizing patients' information in often chaotic, difficult to manage, paper-based systems [11]. The Personal Health Book can partially alleviate these problems by moving some of the responsibility for preservation of basic medical records from institutions to the patient.

This paper introduces the concept of Portable Health Book (PHB) for developing countries. The PHB is a patient-centred software application that enables consumers to collect health information from one health clinic, view the information, and

take the information to another health clinic. The paper introduces a PHB use-case for offline USB Smart Card and online smart phone cloud environment storage.

In the Methodology section, we introduce a PHB usability use-case in a developing country context and simplified interoperability requirements. In the technical design description and user interface sections, we show how to import data from a source health clinic, describe the structure of an adaptable decomposed database, and indicate how the stored data can be viewed. We then test the PHB framework by importing PHRs from s developing country database and Japanese hospital and discuss the results and implications.

II. Methodology

A. Portable Health Book Usability

The implementation of a digital Portable Health Book in developing countries requires affordability (low cost) and simplicity (ease of use). Following these drivers are robustness, privacy, and security. For this study, it is suggested that a USB or Smart Card [12] is a suitable and affordable solution for the low-income person. The "ease of use" depends on the literacy level of the consumer and willingness of the healthcare provider to share the PHR. We envisage that the consumer will ask the healthcare provider to store the resultant PHR or EHR on their USB or Smart Card. This is applicable to rural community health clinics or visiting health providers such as the Grameen Portable Health Clinic in Bangladesh [13]. At the rural community, health clinics store computer records and give patients printed prescriptions. It is feasible to ask for a digital copy. Once imported a USB or Smart Card, the consumer will store the PHR and then can view PHB data using a PC or tablet.

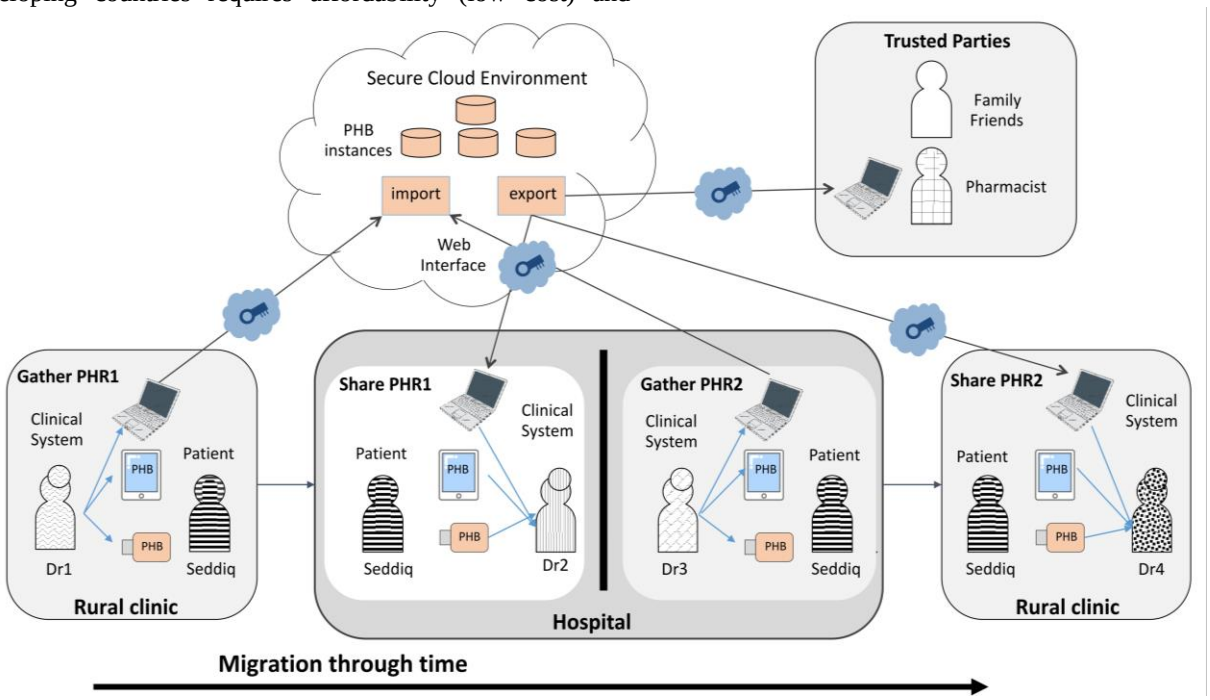


Fig. 1. The Portable Health Book use case

As low literate consumer in developing countries often migrate from rural villages to urban centers, the consumer should be able to ask another healthcare administration or doctor in a hospital to import their latest PHR or health history from the USB or Smart Card they are carrying. This will give the health clinic an improved perspective on the consumer's medical history. Following the consultancy, the patient can ask for the results to be imported to the USB or Smart Card. The PHRs will increment and the consumer will ask the next health clinic to view the appended PHRs and so on.

To provide robustness to the Personal Health Book application requires extending the storage options to the low literate consumer. The ability to save the PHR to an external storage such as a cloud environment will overcome the risk of physically losing the USB/Smart Card or file corruption. In this instance, the consumer will ask the healthcare provider or a family member to export the PHR to a secure website. The consumer will need a username and a password to access their unique remote PHB database.

A cloud-based PHB service has the benefits that the low literate consumer can use a smart phone PHB App to manage and store the PHR both on the smart phones and at the internet site. Once authorized by the consumer (health information owner), other interested parties such as family, friends and ultimately healthcare professionals, such as doctors, can view and import PHRs.

An example use-care for the PHB is shown in Fig. 1. Seddiq is a low literate middle-aged male consumer with hypertension and is examined at the rural clinic in Bangladesh. He would like to import the PHR. He has a choice of USB Smart Card with a PHB application (5 MB); a smart phone with PHB App but he needs a transmission method for importing the file; or asking the healthcare provider to export

the PHR to a website (he has to type in the username and password). Seddiq, later migrates to urban city, overeats and feels unwell. He visits the hospital and allows the administration to access and import his PHB. He chooses the website method and shares his username and password. After the consultation, he asks the administration to export a copy of the PHR to the web-based PHB. Seddiq's family rely on his earnings, they are concerned his health status, and occasionally view a summary of his PHRs. During the festival season, Seddiq returns to the rural village and visits the health clinic for a check-up. Due to Internet outage, the community health clinic import his health records from his USB Smart Card. Using this approach, Seddiq, the people that depend on him, and health providers, have a convenient method to manage the PHR and make important decisions.

B. Personal Health Book and Interoperability

A unique characteristic of the PHB is that it tries to overcome interoperability issues by using a common file formats such as XML. This is out of synchronization with the current interoperability expectation to use HL7 as a common interface between healthcare providers. However, in developing countries, most health clinics have not implemented HL7 and have severe budget constraints. Thus in low resource areas, it is more pertinent to use the existing computers systems, databases and standard interfaces. Accepting this, there is still the problem that each health clinic may have its own data formats and structures. The PHB role is to accept the health clinic file, efficiently store the file, present the file to the PHB owner in a unified view, and create an understandable report to the next health clinic. To achieve this goal there are three main PHB functions:

Gather: PHRs are imported medical records from various medical sources (clinics, hospitals); PHRs contents are read; and stored it in the PHB database.

Store: PHRs will have different data characteristics: data type (integer, char, date); format (size of each data, storing order of data) and data structure (names and number of columns). PHB database will accommodate all incoming data in sufficient and reliable manner. The PHB owner should be able to understand the imported data.

Share: The PHB owner should be able to access, understand, manage and share the data with (clinics, doctors and family) and have full control of it.

C. Privacy and Security

In this paper we are introducing the primary concept for the Personal Health Book. Personal privacy and data security are under examination. PHB data will be securely stored into the PHB databse by encrypting all personal data [14], and no access will be granted to anyone without securely logging in the system [15] by having a user name and password [16]. The PHB can get higher security by incorporating the geomatric approach [17] or the token system [18] during login attempts.

III. PHB Technical Description

Personal Health Book is a software application for carrying a patient's health information as PHRs. The main criterion for the design is application size so it can run on any device including a smartcard chip. We focus on optimizing and simplifying the logic so the application can be implemented in budget-constrained environments.

A. Importing from Source Health Clinic

To import a PHR from a health clinic, the current concept focuses on XML file format and has basic expectations. To achieve a successful import to the PHB database, the following is required:

- File format: XML
- XML structure:
    - Group items by component types
    - Each consecutive component must be nested to its parent
    - For more than one word component name, a component must have "<type>" item with the value of the component's name.
    - Items (Fields Naming): Correct English spelling
- Minimum component types:
    - Patient
    - Hospital/Clinic
    - Physician/Doctor
    - Nursing/Checkup

These set of rules are translated into an XML schema file (XSD) used for XML validations, which also helps in the logic of reading the XML file content.

B. PHB Import Logic

To import the XML file from the source health clinic to the PHB database (see Fig. 2), we designed an algorithm to import PHRs coming from different sources. The steps are as follows:

1) Check imported file extension:
- If it was not an XML, go to step 4.

2) Validate XML file structure against a predefined PHB schema (XSD):
- If it was not valid, go to step 4.

3) Open XML file and read its contents:

a) Get all parent nodes

b) For each parent node, check it's availability in the "Categories" table in the PHB database:
- If it does not exist, create a new category

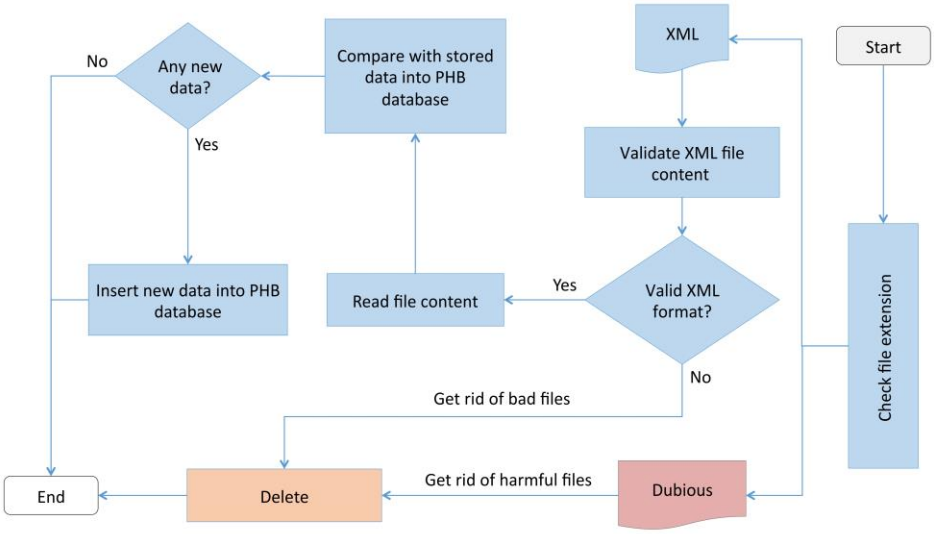


Fig. 2. The PHB Workflow for Gathering and Storing PHRs

c) or each child under each parent, check it's availability in the "Items" table in the PHB database:

- If it does not exist, create a new item

d) For the existing category, check if imported child node match the number and order under existing templates in the PHB database

e) Create new templates for new categories and for the mismatch cases from point d).

f) Insert new values into the "Values" table in the PHB database according to templates ordering of the items

g) Insert new records for new categories and link them to their values and templates

h) Insert related medical records and link them to their sources, values and templates

4) Close the XML file

C. PHB Database

The PHB database needs to manage files and data from multiple healthcare sources; it should adapt any new data type and grow dynamically. It should work with future changes in the data without the need to change the database structure.

The entity relationship data model for an adaptive PHB database is shown in Fig. 3. The ER diagram depicts the interconnections and relationships between entities: User, Record, Source, Category, Value and Item. For example, User and Record are connected via the relationship "Owns" (one-to-many). In other words, the user will own many records. Similarly, entities Source and Record are connected via the relationship "Imported from". Here, a record will be imported from a source. In another scenario, Record and Value are connected via "Has values" . Which means, a record will have

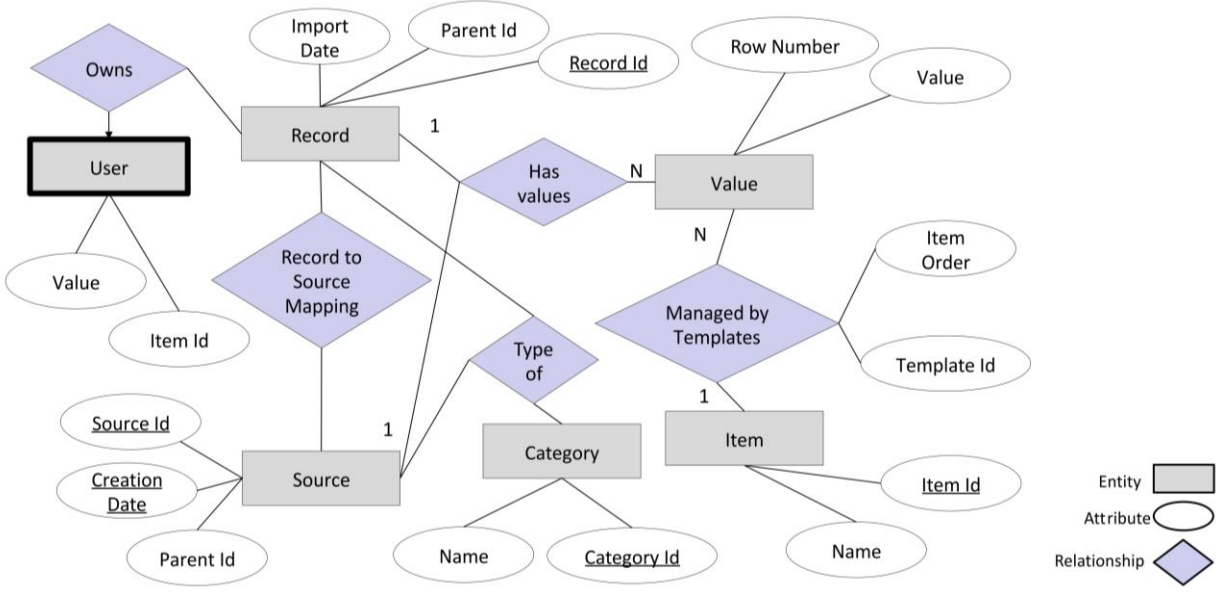


Fig. 3. Entity Relationship Model for the PHB Database

many values. Similarly, other connected entities have relationships in a meaningful way.

a) PHB database design: The PHB database consists of 6 tables. We used MS SQL to create basic database concepts that will work with all database engines. The tables and their functions are described as follows:

- User Profile table: User personal details, such as name, date of birth, and address.
- Items table: All data naming/labels are stored as an index (including unknown new items). An item can be name, age, temperature, height and description.
- Categories table: This table is used to manage data types, such as hospital, doctor, drug and checkup details.
- Templates table: As PHB accepts all data coming from different sources with different structures, there is a need to keep track what items related to each source.
- Records table: All health records are stored in records table as well as related records. For example, one hospital visit record has prescribed drugs and lab reports; these records have their own related data, different elements and more than one occurrence. They will be saved as sub-records in the records table. Even records source details will be kept in this table, such as hospitals, doctors, and drugs. A source can be sub of another source (doctor is sub of hospital). Each record sored in the database, will be related to a certain source, such as, a doctor can be a source of a record, or a drug is the source of the prescription. All sources (Hospital) to sub-sources (Doctor) and records (Checkup details) to sub-records (Prescriptions) are linked through "ParentId" column in the "Records" table.
- Values table: All the values of the items that are specified in the templates table are stored here, which is actually the values of the sources and records details.

The database is decomposed and designed in such a way that there will never be a "Null" value or a blank stored in any table (Fig. 4).

b) Querying the PHB database: Every source and record will have a set of values linked together by the template table. To query the PHB database, the templates table will identify the items (naming/labeling) related to each value of the queried data type (source/record). Items related to the values are stored in the Items table, and every template is linked to a number of these items. If a hospital changes the number of items, it will not affect the existing stored data because the date instance is kept. An example of a query would be to type in a keyword such as 'Blood Pressure' between selected dates and the PHB would present the report.

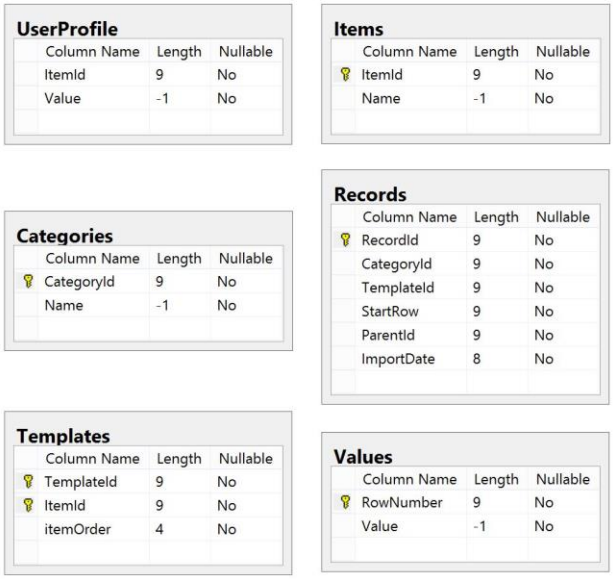


Fig. 4. The PHB Decomposed Database Design

D. Share Stored PHRs with a Doctor from Another Clinic

The PHB has the ability to share the patient's PHRs visually through PHB application directly, so a doctor can understand the patient's health history, in addition a report can be generated and exported to another system.

To export a PHR report from the PHB database to a destination (seeking) health clinic, the current concept will generate an XML. To achieve a successful export that can be understood by the seeker, the following steps will occur:

1) Get all root sources
2) Create new blank XML file and open it to be filled
3) For each root source, get its details (by referring to templates)
a) Get all child sources of the root source
b) For each child source, get its details (by referring to templates)
c) Get all parent records of each child source, get its details (by referring to templates)
d) For each parent record get all sub-records that their sources are sub-sources of the child source, get its details (by referring to templates)
e) Get sub-records that their sources are not sub-source of the root source, get its details (by referring to templates)
f) Fill into the XML file all collected data
4) Close the XML file and publish it

IV. User Interfaces

This section indicates how a consumer can use the PHB when ported onto a USB card. Whenever the patient visits a hospital, the following steps occur:

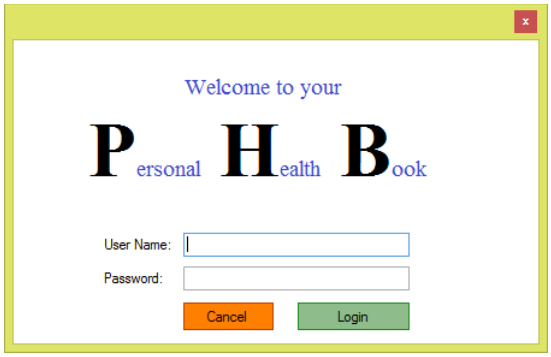


Fig. 5. Log in screen

1) The pateint gives the USB smart card to the doctor or health-worker to plug it in their personal computer.

2) The doctor/health-worker runs the PHB application, and a login screen will appear.

3) The patient has to insert his username and password as shown in Fig. 5.

4) After a successful login, the PHB's main window will appear, and the user can navigate through the application via the main menu (see Fig. 6).

The menue has three main items:

- Reports: Where patient can view and share his health related reports (Health Reports) or understand his data statiscally (Statistical Reports)
- Import PHR: Where doctor/healthcare worker can import new PHRs
- Logout: It is important that the patient logs out after he finishes from the PHB application to keep his data securely safe.

5) If (Health Reports) was selected, health reports form will appear (see Fig. 7), here the patient can view his health related data.

6) The (Generate Report) button will produce the health report based on the patient's customisation and filtering.

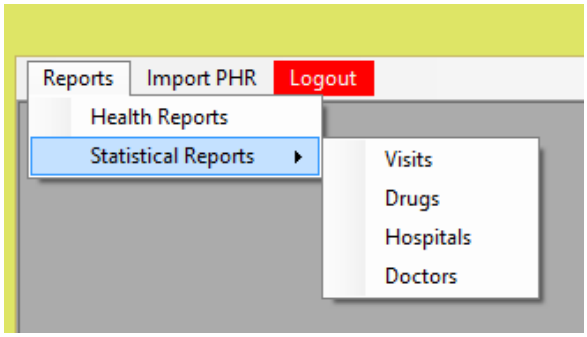


Fig. 6. The PHB Application Main Menu

7) By clicking on the (Export) button, the patient can share a copy of the produced report with the doctor/health-worker, either in an XML or PDF format. As a secuerity and privacy protection measure, an authenticaion form will appear and the patient will have to enter his credintials to allow the export process.

8) After the patient finishes the health checkup, the doctor/health-worker will be able to import the new PHR to the PHB application by selecting (Import PHR) from the main menu.

9) To quit the application, the patient simply clicks on (Logout).

V. Testing the PHB

A. Importing real PHRs and viewing the output

To test the PHB framework we randomly selected a real patient PHR (with personal details deleted) from the Grameen Portable Health Clinic database in Bangladesh (22,000 entries). We accessed 2 PHR records from a Japanese hospital "Hospital xyz" (representing an urban city) issued at different times. We imported 3 PHRs into the PHB environment. The PHB role was to reorganize the data from two health clinic visits by categorizing and grouping and presenting a unified, human understandable, view structure.

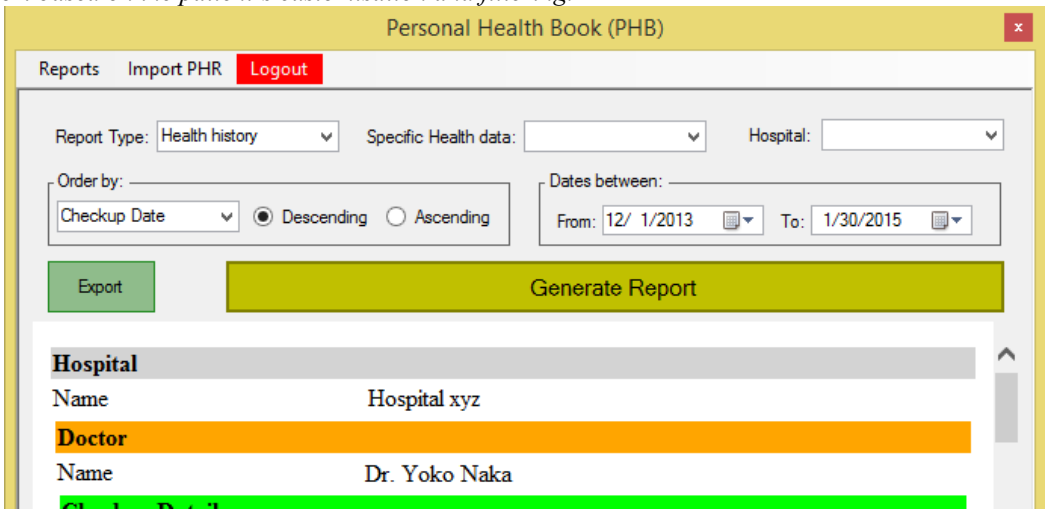


Fig. 7. Health Reports form

The Portable Health Clinic normally generates PDF files to print PHRs. For the test, we created an XML file to match the import requirements (see Fig. 8). The data was successfully imported in the PHB database.

We implemented an experimental version of the PHB (USB card version) to test the readability and the capability of the PHB application. In which the consumer can export generated report as a PDF of XML file to share it with his doctor, family or hospital. Because the PHB database is designed to get the use of each entity stored, it can produce health related reports as well as statistical reports such as number of visits, drugs, hospitals and doctors. These types of reports can help in understanding the extent a person is following up on his health and the address and contacts for all visited hospitals and doctors.

The PHB patient report is human readable, understandable and meaningful as seen in Fig. 9. The reports of multiple visits from two medical firms are shown in a descending order. Every category/type in the report is clearly separated from the others and contains the exact number of items that the original record contained. In addition, all child's components are displayed under their parent source.

B. Testing the impact of many health clinic visits

To assess the impact of visiting many health clinics and subsequent creation of PHRs on the PHB application data size, we prepared a simulation to observe the rate of size growth in bytes every time a health record is imported (see Fig. 10). This is important because we do not want the PHB application to grow and exceed the memory size of the USB card. The simulation involved a patient visiting 2 hospitals a total of 24 times and generating 24 PHRs. The name of doctor was randomly changed and whether drugs were prescribed or not was randomly selected. The simulation output showed that the first imported PHR into the system consumed a large memory

```
<?xml version="1.0" encoding="utf-8" standalone="yes"?>
<content>
  <Source>
    <Type>Hospital</Type>
    <Column1>Name</Column1>
    <Column2>Portable Health Clinic</Column2>
    <SubSource>
      <Type>Doctor</Type>
      <Column1>Name</Column1>
      <Column2>Dr. Ahmed</Column2>
      <Column1>Address</Column1>
      <Column2>Dhaka, Bangladesh</Column2>
      <Checkup_Details>
        <Type>Checkup Details</Type>
        <Column1>Checkup date</Column1>
        <Column2>01/12/2013 11:04 AM</Column2>
        <Column1>height</Column1>
        <Column2>153</Column2>
        <Column1>weight</Column1>
        <Column2>53.5</Column2>
        <Column1>bmi</Column1>
        <Column2>22.85</Column2>
        <Column1>waist</Column1>
        <Column2>89.3</Column2>
        <Column1>hip</Column1>
        <Column2>91.6</Column2>
        <Column1>waist_hip_ratio</Column1>
        <Column2>0.97</Column2>
        <Column1>temperature</Column1>
        <Column2>98.67</Column2>
        <Column1>oxygen_of_blood</Column1>
        <Column2>98.6</Column2>
        <Column1>bp_sys</Column1>
```

Fig. 8. Imported EHR from PHC database, Bangladesh

Hospital	Last Visit
Name	Hospital xyz
Doctor	
Name	Dr. Yoko Naka
Checkup Details	
Checkup date	24/01/2015 10:00 AM
Metabolic	No
Medical History	No
Advice on Treatment	No
Blood Pressure	84/52
⋮	
Hospital	Older Visit
Name	Portable Health Clinic
Doctor	
Name	Dr. Ahmad
Address	Dhaka, Bangladesh
Checkup Details	
Checkup date	08/08/2014 10:00 AM
weight	53.5
BMI	22.85
waist	89.3
hip	91.6
waist_hip_ratio	0.97
⋮	
Hospital	Older Visit
Name	Hospital xyz
Doctor	
Name	Dr. sanji Jiro
Checkup Details	
Checkup date	21/06/2014 10:30 AM
Chest	84.0
BMI	23.1
Fat level	+ 5.0 %
Blood Pressure	102/60
⋮	
Doctor	Oldest Visit
Name	Dr. Yoko Naka
Checkup Details	
Checkup date	30/12/2013 11:04 AM
Blood Pressure	120/72
Body Measurements	
Height	167.4

Fig. 9. PHB Personal History Report

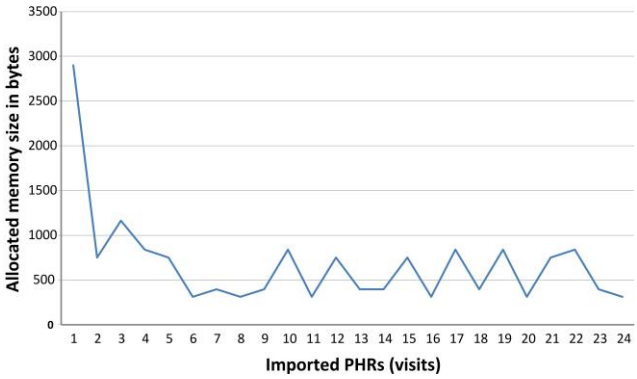


Fig. 10. Database Size Increment vs. Number of Imported Records

size as new items, categories and templates were created by the PHB database. The following imported PHRs consumed much less memory size as the database structure had been set up. When new items/categories are created they consume memory but less than the initial set up. Therefore, Fig. 10 shows a drop in bytes after the first PHR and then moves to a steady state as a few new items/categories are introduced. When the consumer visits the same hospital with same doctor and receives the same medication there will be minimal impact to PHB memory consumption (PHR number 7, 9, 13, 14 and 23).

VI. Results & Discussion

The PHB test results demonstrate that PHR XML files can be imported, the patients' characteristics and medical data stored in a decomposed database and human readable reports can be generated. A subsequent XML file imported to the PHB, such as the next medical check-up in a different source health clinic, will append the data and appear above the existing data. A healthcare worker at another health clinic reading the report can export a PDF or XML copy to their system or print the report and manually update their system. Following the health consultancy, the patient will ask the healthcare worker at the clinic for a digital report containing the results to be imported into the PHB.

The impact of visiting different health clinics over time on the PHB memory requirements is minimum. The decomposed database can grow dynamically as new data types are added (the customer has a new disease or health problem) and is designed that there will be no null or blank values stored in a table leading to optimized memory management. The simulation indicated an average of 500 bytes per PHR per visit. With this indication, we can assume that an average patient can store a lifetime of personal health history on a USB Card.

There are many challenges for the widespread adoption of a PHB in developing countries. Health clinics in the rural areas may not have IT systems and cannot give digital records (XML files) to their patients. The source health clinic may not want to share the patient's health records with the patient. The seeking health clinic may not want to accept an unknown patient's PHB in fear of a computer virus. There may be affordability and literacy issues with USB sticks or accessing the internet. There are security and privacy concerns and many technical questions. However, a PHB implementation would introduce a robust patient-centred healthcare model and a unique opportunity for low-income individuals to manage their own health in many countries where healthcare systems are under severe pressure.

VII. Conclusion

Personal Health Book software is an application that manages Personal Health Records and is suitable to developing countries where low-literate consumers migrate from rural areas to urban areas and who want to know their medical history. It provides free aid to low resourced healthcare institutions facing difficulties in implementing Electronic Health Systems by allowing the institutions to save standard files in their existing databases. By porting the Personal Health Book application onto devices such as USB, smartcard or low-cost smartphone with connection to cloud computing environment, will enable low literate consumers to access their own medical records even if they migrate to another country.

The PHB application is a single-user database smart health convenience option for the consumers rather than a medical necessary. A PHB cloud-based solution is an aggregate of many single personal unique databases rather than a large common database. There are many risks to storing all PHRs in one location, such as privacy, theft, and non-acceptance by health professionals and this is beginning of the list. However, like most tools of convenience it will take time and modification before becoming generally adopted. The most important PHB attribute is that allows individuals to easily store, access and review their own personal digitized health history. The PHB is thus more relevant to the individual than to the institution that may use it to make clinical decisions. The PHB is also relevant for individuals who are increasingly becoming interested in quantifying their lifestyles and are using sensors to measure and make sense of their wellness over time.

VIII. Future Works

Personal Health Book application requires more development and testing. The application requires robust security measures to protect the consumer's private health data using a strong authentication and encryption method and against unwanted external intrusions. Error and correction management to handle file corruption during PHR import process. The cloud environment architecture will be defined. Testing involves importing different types of PHRs from health clinics and their departments. Checking if health clinics can import a real PHR stored on the PHB to their data system and researching the associated problems. Finally, surveying end-users - low literate consumers and medical professionals – on the applicability of the PHB to their lives.

Acknowledgment

We thank Global Communication Center, Bangladesh, for access to and use of personal health records from Portable Health Clinic database.

References

[1] Claude Sicotte, Jean Louis Denis, Pascale Lehoux and Francois Champagne, "The computer based patient record challenges towards timeless and spaceless medical practice", Journal of Medical Systems, vol 22, no.4, pp.237-256, 1998.

[2] Marion J. Ball, Melinda Y. Costin and Christoph Lehman, "The personal health record: Consumers banking in their health", Stud. Health Technol. Informat., vol 134, pp.35-46, 2008.

[3] Robert Steinbrook, "Health care and Amercian Recovery and Reinvestment Act" N.Engl. J.Med., vol 360, no 11, pp.1057-1060, Mar 12, 2009

[4] Maisie Wang, Christopher Lau, Fredrick A. Matsen and Yonfmin Kim, "Personal Health Information Management System and its Application in Referral Management", IEEE Trans on Information Technology in Biomedicine Vol.8., No. 3. September 2004.

[5] Healthcare Information Management and Systems Society (HIMSS). (2007). HIMSS personal health records definition and position statement. Retrieved October, 22, 2009, from http://www.himss.org/content/files/phrdefinition071707.pdf

[6] Kim Kyungsook and Eun-shim Nahm, "Benefits of and barriers to the user of personal health records (PHR) for health management among adults", Online Journal of Nursing Informatics (OJNI), 16 (3). http://ojni.org/issues/?p=1995

[7] Kim Nazi, "The Personal Health Record Paradox: Health Care Professionals' Perspectives and the Information Ecology of Personal Health Record Systems in Organizational and Clinical Settings", Journal of Medical Internet Research. Vol 15 No 4. April 2013. http://www.jmir.org/2013/4/e70/

[8] Juha Puustjarvvi and Leena Puustjarvi, "Personal Health Book: A Novel Tool for Patient Centered Healthcare", In the Proc of the International Conferenece on Health Informatics (HEALTHINF 2011).pp 386-393, 2011

[9] Juha Puustjarvi and Leena Puustjarvi "Personal Health Book: A Cloud-Based Tool for Patient Centered Healthcare", Journal of Public Health Frontier. Vol 2., Iss 3., pp 146-155, Sept 2013.

[10] Seddiq Alabbasi, Andrew Rebeiro-Hargrave, Ashir Ahmed, Kazuaki Murakami, and Hirohito Yasuura, "Personal lifetime data and its smart management" Proceedings of the 2013 International Symposium on Intelligent System Engineering (ISISE 2013), November 15-16, Abu Dhabi, UAE

[11] Weihua Chen and Metin Akay, "Developing EMRs in Developing Countries", IEEE Transactions of Informationa technology in Biomedicine, Vol 15, No. 1 January 2011

[12] Francine L. Maloney and Adam Wright, "USB-based personal health records: An analysis of features and functionality", Int. J. Med Informat", vol 78, no.2, pp97-111, Feb 2010

[13] A. Ahmed, A. Rebeiro-Hargrave, N. Yohara, E. Kai, R. Hossein, N. Nakashima "Targeting Morbidity in Unreached Communities Using Portable Health Clinic System". IEICE Transactions on Communications, Vol.E97-B, No.3. March 2014

[14] Rajitha Tennekoona, Janaka Wijekoona, Erwin Harahapa, Hiroaki Nishib, "Per-hop Data Encryption Protocol for Transmitting Data Securely Over Public Networks", Procedia Computer Science, Vol 32, pp. 965–972, 2014

[15] Stuart P. Goringa, Joseph R. Rabaiottib, Antonia J. Jonesb, "Anti-keylogging measures for secure Internet login: An example of the law of unintended consequences", Computers & Security Vol 26, Issue 6, pp. 421–426, September 2007

[16] Khosrow Dehnad, "A simple way of improving the login security", Coputers & Secruity vol 8, Issue 7, pp. 607-611, Novemeber 1989

[17] Tzong-Chen Wu, "Remote login authentication scheme based on a geometric approach", Computer Communications Vol 18, Issue 12, pp. 959–963, December 1995

[18] Yen Sung-Ming, Liao Kuo-Hong, "Shared authentication token secure against replay and weak key attacks", Information Processing Letters, Vol 62, Issue 2, pp. 77–80, 28 April 1997

En-Route Vehicular Traffic Optimization

Saravanan M
Ericsson Research India
Ericsson India Global Services Pvt. Ltd
Chennai, India

Ashwin Kumar M
Dept. of Electronics and Communication Engineering
Meenakshi Sundarajan Engineering College
Chennai, India

Abstract*—The pathways of information are changing, the physical world itself is becoming a type of information system. In what's called the Internet of Things (IoT), sensors and actuators embedded in physical objects—from roadways to pacemakers—are linked through wired and wireless networks, often using the same Internet Protocol (IP) that connects the Internet. When objects can both sense the environment and communicate, they become tools for understanding complexity and responding to it swiftly. The revolutionary part in all this is that these physical information systems are now beginning to be deployed, and some of them even work largely without human intervention. This paper has addressed the traffic congestion problem with the help of Internet of Things. Increase in the number of vehicles in cities caused by the population and development of economy, has stimulated traffic congestion problems. It is becoming more serious day after day in the present scenario of developing countries. The reason for the same could be categorized as mismanagement of vehicular movement, ineffective system for controlling the mobility of vehicles, uneven roads and traffic snarl-up. Unexpected vehicular queuing is a major concern leading to wasting time of passengers and thwarting ambulance to reach the destination in time. In addition to that, traffic congestion makes it difficult to forecast the travel time accurately causing drivers to allocate more time in travel than scheduled previously. To ease these mounting traffic problems a demonstration is made on the Proof of Concept (POC) using the smart city data set provided by Telecom Italia of Milan city, to verify that these concepts have the potential for real world application and could be used by the government sectors or private transport organizations to ameliorate the passenger's comfort on road which are as follows. A central node is developed which sets the speed limit and predicts a normalized speed separately for each locality from the available data set. For efficient control in mobility of vehicles an advanced dynamic digital board is introduced, which displays the speed limit set by the central node time to time. The normalized speed could be used to estimate the effective time taken between destinations precisely. By comparing normalized speed with real time values anomalies in the locality like congestion and presence of uneven roads is predicted. Accident detection model is integrated with the central node which sends a message to dynamic board indicating location of the accident along with the time taken. It even improves traffic flow around the accident occurred location. Central node together with navigation tools could provide re-routed path to the drivers during congestion or accident.

Keywords—IoT; IP; POC; Central Node; Dynamic Board; Accident detection model

I. INTRODUCTION

Machine to Machine (M2M) communication along with its relevance IoT (Internet of Things) technology is simple, embedded and invisible to connect billions of devices to introduce Networked Society [1]. Many cities and homes are changed to the new norm of smart in the sense that all the devices attached are interconnected and communicated with each other for the benefit of the society uplifting. The connected device details of one of a smart city are analyzed and reported in this paper.

Traffic congestion is a major concern to all countries, be it, developing or developed. Unexpected vehicular queuing is a major concern leading to wastage of time of passengers and thwarting of ambulance to reach the destination in time. Adding to that, traffic congestion makes it difficult to forecast the travel time accurately causing drivers to spend more time in travel than scheduled previously. The idea is to optimize traffic congestion solutions with the help of existing intelligent sensor devices. These devices can measure many parameters for a more efficient route management of the city. The remote sensor gathers data and sends it wirelessly to a network, where it is next routed, often through the Internet (with IoT facility), to a server such as a designated node. At that point, the data is continuously recorded, analyzed and acted upon the connected devices with less human intervention. The block diagram (Fig 1) shows the integrated view of various components built in the proposed system. Here central node acts like a hub in receiving messages from sensors and vehicles during an accident and communicates this with the dynamic board by sending the message it has to display. The main contributions of this work are to address the real issues pertaining to vehicular traffic optimization in a smart city.

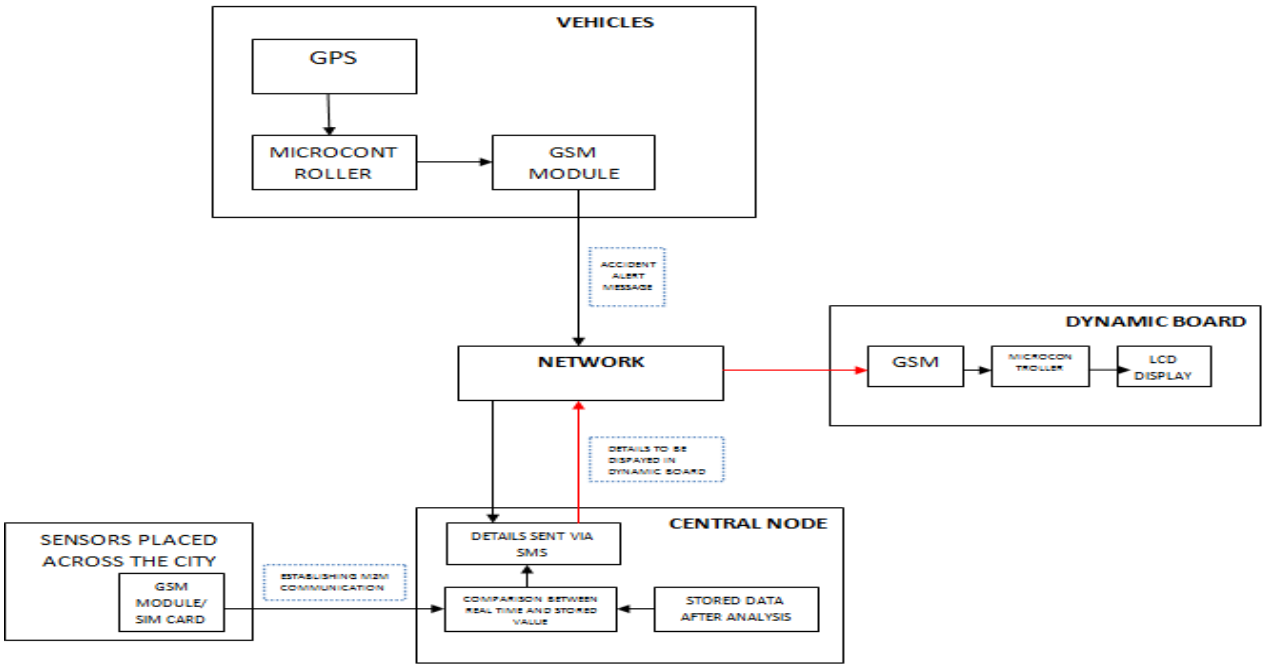


Fig. 1. **Block diagram.** Central node acts like a hub receiving messages from sensors and alert messages directly from the vehicles during accident and communicates with the dynamic board by sending messages it has to display. GSM is used for sending and receiving messages, black and red arrows are used to differentiate receiving and sending messages respectively

A. *Management of vehicular movement*

The central node is introduced, which predicts the speed limit across the city in an hourly basis considering people's movement and traffic flow in a locality and instantaneously varying the speed limit during accidents and unexpected traffic snarl up. Dynamic board is constructed and setup in various places across the city which displays the speed limit predicted by the central node, thereby ensuring efficient control in mobility of vehicles. Along with speed limit, central node predicts normalized speed (median speed) separately for each locality from the available data set. This normalized speed is used to estimate time taken between destinations which help commuters to schedule their travel plan more efficiently.

B. *Traffic snarl-up*

The reasons for traffic snarl-up are unexpected congestion and also the occurrence of accidents. In order to prevent unexpected congestion the rerouted path is provided to the user with the help of navigation tools and dynamic boards. In order to reduce the traffic during accidents in a particular area, accident detection model is integrated with the central node which alerts the nearby hospitals immediately and sends a message to dynamic board indicating location of the accident along with the time taken for ambulance to cross the board through Global System for Mobile (GSM) module. This would avoid congestion around the accident location and also ensure ambulance to reach the destination in time.

C. *Uneven road*

By comparing normalized speed with real time inductive loop (a device laid on the road within certain range) values, any anomalies in the locality like congestion and presence of uneven roads could be predicted by understanding the change in speed of the vehicles.

II. Related Work

Traffic congestion problem had numerous solutions in the past. There have been multiple systems which address specific parts of the problem but lacked to provide end-to-end solutions and faced many difficulties to prove its effectiveness. Here discussions about few such systems along with its limitations are made. The existing works of predicting the speed limit is based on peak hour traffic and weather changes [2]. The concept of offline routing exists. A particular node speed limit value is set only based on the available data of peak hour traffic, static speed limit and change to another static value during the presence of rain or other reasons. Its occurrence is detected with the help of rain sensors equipped in speed board. This leads to frustration of travelers when they have to abide to the speed limit set based on peak hour traffic during night hours. In addition to this, fixed speed limit leads to lack of significant control over mobility of vehicles. Even in some developed countries, variable speed boards have been designed with only few parameters being taken into considerations like lightning and construction zone [3]. This shows there is a lack of efficient method to set the speed limit, which ultimately leads to mismanagement of vehicular movements.

Another system includes a Global Positioning System (GPS) based navigational system that includes a GPS receiver connected to a wireless communication device for communicating with a remote computer over a wireless communication network [4]. The GPS-based navigation system continuously determines the motored vehicle's exact

physical location in a region that is intermittently or continuously uploaded to a remote computer via the wireless communication network, this work aids at finding the traffic congestion in a given area. But the problem is during such huge traffic the remote computer faces certain difficulties to manage the details of the each GPS receiver. Monitoring such a huge traffic of data is very expensive.

Various sensors such as 3D laser scanning devices along with 3D reconstruction algorithms are used to measure the size of road potholes [5, 6, 7]. In [8, 9], cameras are installed on vehicles to record road videos, from which road conditions are inferred. With the help of Accelerometer, vibrations on the road can be effectively captured from the vehicles and along with GPS, its actual location is determined [10]. The system to detect the presence of uneven roads uses 3D laser scanning devices or cameras to record images and videos or Accelerometer plus GPS equipped inside a car, but all these three systems are too expensive for wide adoption and communication from a car to a remote control stills remains a challenging task as mentioned earlier.

Some research studies were carried out on accident detection system, aerial surveillance or close circuit surveillance [11]. All these systems are used to determine accident but the drawback of this system is that, someone has to witness and report the accident. When severity of the accident is high, the driver reporting accident system fails, as the driver may not be able to report in such conditions. Smart phone based accident detection models [12] are very expensive and even lead to false alarm. Continuous monitoring of vehicular GPS device through a remote computer help to detect accident instantaneously but monitoring the huge traffic of data is challenging [4]. Researchers have also reduced the traffic by sending only alarm to remote computer during accident occurrence with the help of microcontroller, GPS and GSM equipped vehicles [13] but still it lacks end-to-end results because it lacks efficient rescue operations.

Another typical problem is to predict the accurate travelling time in advance which is mostly required for the specific industry need and people's daily life. Researchers have come up with some systems for estimating efficient time taken. Travel time prediction algorithms attempt to estimate the time of travel between an arbitrary origin-destination pair in a roadway network, for current and future instances. It assists the driver to choose a less congested route, thus can be used for optimal routing and dynamic route guidance [14, 15]. There is also a model to estimate the travel time that begins at a long-term future moment of departure [16]. This system lacks real estimation of time, because several factors like current traffic, congestion due to occurrence of big events, presence of accident and so on are not taken into considerations. In this paper, we are addressing these challenges to bring out a better system to optimize and improve the present smart city environment.

III. Dataset Description

At present, data is woven into every sector and function in the global economy. The use of Big Data Analytics — large pools of data that can be brought together and analyzed to discern patterns to make better decisions in enhancing productivity, creating significant value for the world economy by reducing waste, increasing the quality of products and thereby improving customer services. Telecom Italia on January 2014 released different categories of data like telecommunication, energy, weather, health insurance company details, public and private transport, Social Networks and events to the international scientific community. The data provided within the dataset of the Big Data Challenge will be geo-referenced and anonymous relating to the territories of Milan and of the Autonomous Province of Trento [17]. The dataset contains millions of records of data covering the period from November to December 2013.

By utilizing the dataset provided by Telecom Italia [17] and by categorizing city details into referred grids (each square id of area 235*235m^2), the people movement in a grid is estimated with the help of available call detail records. For predicting median speed (normalized speed) in a grid, the inductive loop detectors (sensors installed across city providing details of number of cars crossing the detector and speed of each car along with its direction) are used. The private insurance company has installed GPS device inside a car for its customers. Using this, the traffic data is obtained and it also provides details of number of users in each grid along with the average speed.

IV. System Overview

The main components of the system include central node, dynamic board and accident detection system in vehicles. Available data are processed and stored in central node and real time communication between central node and various sensors like inductive loop detector, precipitation sensor and vehicles-accident detection model is established. With the help of GSM module, central node communicates with the digital board via Short Message Service (SMS) regarding details it must display. These digital boards are called as *dynamic board* in this paper, as it displays the streaming of information on a continuous basis. Detailed explanation of each of the attached component is given below.

A. Dynamic Board

Dynamic boards are used only to receive text from central node and display the speed limit, location of accident and time taken for the ambulance to cross. Each board composed of Liquid Crystal Display (LCD) screen to display the message, GSM module for receiving messages and a microcontroller which acts like an interface between the two.

1) GSM module - It is used to receive the information (SMS) from the central node. The information received from the central node is displayed in the board. Normally speed limit will be sent from central node and incase of accident, its location and time taken for the ambulance to cross the dynamic board will be provided.

2) Microcontroller - It acts like an interface between GSM module with LCD Display, it is coded to receive the information from a particular mobile number (central node) and display it in the central node.

3) LCD Display - This electronic display is used as a communication medium between the central node and regular commuters of the route. Dynamic board is just a device which receives details and displays it in the display.

B. Central Node

An Intellect Central Node is a remote monitoring system used for storing the processed data. In addition to that, it receives information from all the available sensors. It compares the real time values along with the predicted values to find if any anomalies are present and during the case of anomalies it takes the necessary steps needed to control the situation. In this study, anomalies are considered as the occurrence of an accident, reduced speed limit in a grid (uneven roads), drastic increment in call detail record value (increase in people movement) and presence of rain (weather sensor). Central node sends the details it should display in an hour interval, but recognising any anomalies in a grid it immediately sends new details, board must display and intimate necessary details along with it.

C. Vehicles

Vehicles act as an accident detection device, it makes use of GPS receiver, Microcontroller and GSM module while GPS receiver tracks all the satellites and provides accurate positioning data in NMEA (National Marine Electronics Association) standards, GPRMC (Recommendation minimum data for GPS) sentences (which recommends minimum specific GPS/Transit data) in every second [17]. In Microcontroller, memory spaces are allocated for speed and position (latitude and longitude), and values are separated and stored by interpreting the position in GPRMC sentences with the help of the number of commas. The Microcontroller unit checks the variation in speed for every second, by comparing it with the previous value. If the variation is more than or equal to 28.8kph for 1sec then, the microcontroller expects the traveller to respond if it's a false detection and waits for 5 seconds. If the traveler fails to respond then accident alert is triggered and sent to central node using GSM module. Speed variation concept (variation of speed for detecting accident) is explained under Section 7.

V. Tools Overview and Activity

MATLAB [18] and LABVIEW [19] software tools were used to develop the simulated model and a brief description about the activities are given here.

A. Tool- based Activities

For the year 2009-2013, vehicles in operation (per 1000) in Italy was 682 and the numbers are increasing day-by-day, it creates a scenario of monitoring or controlling the vehicle movements which as become a strenuous job. In our implementation, all the discussed difficulties and design activities for an efficient management and controlling of vehicle movements are considered. For each activity, concerned components usage is listed out and brief explanation is provided along with its effectiveness for the requirements.

1) Display in Dynamic speed board: Dynamic boards are set up across the city to display speed limit dynamically which in turn help in restricting the speed of vehicle movements and hence controlling the mobility of vehicles in operation. The idea of dynamic board replacing the normal static board is to enhance its effectiveness. That is in static boards speed limit remains the same during the peak hours and midnight resulting in frustration of the travelers. This is overcome by predicting the speed limit based on traffic flow and people movement from the available data and displaying the same. In addition to it, the flow of the vehicles is controlled during abnormalities like accident and sudden increase in people movement in specific location by normalizing the display in speed limit (display dynamically) and thus ensure proper movement of vehicles during crisis. Since the speed limit is set based on the model learned from the previous year data in a grid, accuracies of its prediction remains high and law enforcement could be tightened.

2) Rerouting: Occurrence of accident leads to traffic snarl-up. The best way to prevent this would be by rerouting and minimizing the vehicle movements. With the available navigation tool the concept of rerouting is included. The predicted shortest path for the user is based on the time taken between places. The travel time duration for vehicles is optimized during accidents. Instead of normalized speed, reduced speed limit is taken into consideration. This allows travel time duration to increase considerably and thus the system avoids the particular node (accident location) and provides the shortest path by considering the travel duration.

3) Congestion avoidance: Avoidance during congenstion is managing the flow of vehicles during abnormalities like accidents, sudden increase of people movement in certain location. When central node is alerted about the same with the help of latitude and longitude provided by the GPS inside a car that predicts the accident grid and also its neighboring grids and sends the optimized speed limit message to the same along with its accident occurred location. This provides knowledge about the congested area to the travelers well in advance and would preferably guide them to avoid the route if possible.

4) Detecting uneven roads: Presence of uneven roads is always hindrance to the travelers and installation of sensors for complete monitoring of all roads will not be possible. Here we make use of the available sensors and design a cost effective model. The main idea behind this model is presence of uneven roads would lead to reduced speed of vehicles. By understanding the current real time situation and map it along with the available data of previous year to predict uneven roads. Possible situations could be changed in weather patterns, people movement and traffic flow during government holidays and normal days and so on.

VI. Data Analysis

The analysis of the telecom Italia of Milan city data constitutes integral part of this model. The telecom data are being grouped into seven types which are listed in Table 1. Based on weather phenomenon the predicted value is being stored in modules after performing mathematical

calculations. In each module speed limit, normalized speed, traffic flow and people movement are predicted and stored for each hour separately indicated in Table 2, for all the Square ids (City of Milan is sub divided into 10,000 square ids with area 235*235m^2, in this paper we consider each square id has different locality). Module name being the square id number followed by R (rain) or A (absent) or I (immediate) indicating the weather conditions (i.e.) 2260A1 or 2260I1 or 2260R1.

The available data are being categorized under the following category

- Tuesday to Thursday (type 1)
- Friday (type 2)
- Saturday (type 3)
- Sunday (type 4)
- Monday(type 5)
- Unexpected increase in traffic (type 6)
- Unexpected increase in people movement (type 7)

In this section, few relevant variables have been predicted. This could be useful to implement the activities required to address the traffic congestion problems.

TABLE I. GROUPING OF DATA MODULE BASED ON 7 TYPES

Types	Z=0	Z=1		Z=2(Imm after rain/snow)
		Rain	Snow	
Type 1	2233A1	2233R1	2233S1	2233I1
Type 2	2233A2	2233R2	2233S2	2233I2
Type 3	2233A3	2233R3	2233S3	2233I3
Type 4	2233A4	2233R4	2233S4	2233I4
Type 5	2233A5	2233R5	2233S5	2233I5
Type 6	2233A6	2233R6	2233S6	2233I6
Type 7	2233A7	2233R7	2233S7	2233I7

A. Normalized Speed

The available speed values for different vehicles are stored based on weather conditions and extreme values are eliminated from the available set. The median value of the available set for the particular time interval is taken as a normalized speed. Inductive loop sensor and private insurance company details are being used for predicting normalized speed.

B. People Movement

Number of people in a locality is estimated based on CDR(call detail record) details. SMS IN, SMS OUT, CALL IN, CALL OUT and DATA PACK values are grouped together to estimate the approximate estimate of people in a locality. The movement of people is estimated through relative ranking mechanism (i.e.) highest number of CDR's generated in a certain location for specific time period will be considered as highest values (for ex. 100) and the relative percentage is worked out for the remaining locations related to each square id.

C. Traffic Flow

The traffic flows are being grouped under types shown in Table 1 based on weather conditions and saved as module wise. The traffic flow strength in each square id is estimated similarly to people movement which is based on relativity.

D. Speed limit

The details from Italian Highway Code (Codice della Strada) of speed limit for several categories of roads like urban roads, motorways and so on are extracted. The available data is sandwiched along with the speed limit details to predict the node speed limit based on the traffic flow and people movement.

TABLE II. STORED REPRESENTATIONS OF VARIOUS PARAMETERS

Time Interval	Normalized Speed	Traffic Flow	People Movement	Speed Limit
00:00-01:00	110 km/hr	32	104	130 km/hr
01:00-02:00	90 km/hr	23	167	110 km/hr
:	:	:	:	:
:	:	:	:	:
22:00-23:00	60 km/hr	16	123	90 km/hr
23:00-00:00	90 km/hr	13	157	110 km/hr

VII. IMPLEMENTATION OVERVIEW

Detailed explanation for implementing the activities in MATLAB [18] and LABVIEW [19] is provided in this section.

A. Effective Time Estimation

In the existing time estimation models, approximated speed value is set for a long range of distance and variation of that traffic is not regularly updated. Both the problems are solved by breaking long range of distance into several short range distance groups based on the number of bends. The speed value is provided for each short distance range via Inductive loop detector (if loop sensor falls in the path) or private insurance company details. The average speed value is derived for the particular grid id and the time taken is estimated for each time interval. The node's latitude and longitude are used to estimate the distance value for short range distance groups separately with the help of Haversine formula [20]. Similarly nodes of four different routes are considered for developing a sample map considering small part of Milan, all the edges and junctions in route are set as nodes. The path between two nodes is considered as distance pair and the distance is estimated via Haversine formula. For each distance pair the details about their corresponding grid id is stored. These nodes are assigned names based on the location and with the help of MATLAB, Dijkstra's algorithm [18] is run to find the shortest path between the destinations. Since no negative edges are being used, the

Dijkstra's algorithm is ideal to find the shortest path from the sample map.

$$a = \sin\frac{\nabla\varphi}{2}\sin\frac{\nabla\varphi}{2} + \cos\varphi 1\cos\varphi 2\sin\frac{\nabla\aleph}{2}\sin\frac{\nabla\aleph}{2} \quad (1)$$

$$c = 2.a.\tan 2(\sqrt{a}, \sqrt{(1-a)}) \quad (2)$$

$$d = R.c \quad (3)$$

where φ is latitude, $\aleph$ is longitude, R is earth's surface (mean radius = 6371km

B. Accident Detection

Accident is detected through vehicle movements by monitoring the speed variation. It is predicted that if the speed variation is equal to or greater than 28.8km/hr for a second, then there might be an accident or abrupt halt of vehicles irrespective of its speed. The two different initial speeds vary between 90km/hr and 110km/hr. The respective final velocities are 61.2km/hr and 81.2km/hr. which is derived with the help of the elementary physics equation

$$v = u - a*t \quad (4)$$

It is understood that the difference between the initial and final velocities as 28.8km/hr in both the cases (90-61.2, 110-81.2).

C. Dynamic speed board

Dynamic speed board utilises the data related to speed limits stored in central node. These boards are installed at various locations across the city should be used as a device to communicate. In this study, the Milano grid number is used as the reference number for the dynamic board, as explained earlier the data is processed separately for each locality (for each square id) and central node is pre-programmed to send speed limit for every hour interval. For efficient control on the mobility of vehicles, the dynamic boards must be set up in the various junctions and lanes where the vehicle movements are high. Since the data provided to us are inductive loop detector values, it could be more advantageous and effective if dynamic boards are set up near these detectors. The weather phenomenon is unpredictable, weather monitoring system in a city immediately send a message to central node about the same and central node sends a message about the new speed limit immediately. GSM receiver is used in the dynamic board which receives the speed limit as a message and protocols are written which reads and writes the message (speed limit) in the LCD display.

D. Traffic congestion avoidance

During the time of accident, the GPS coordinates are received from accident detection module inbuilt in vehicles and based on the available latitude and longitude the grid id is predicted. The nearby grid id is chosen and the alert message is sent from the central node to the dynamic board indicating the change in speed limit, alerting the traveler the location of the accident. This helps local people to change the route on their own which would avoid congestion near the accident location. The shortest time taken path between the hospital and accident location is predicted and time estimation model is used. The source is from nearby hospital and destination is each dynamic board in the ambulance route (location of both is already stored in the central node). The time taken for ambulance to cross the dynamic board is estimated and it is sent as a message along with the speed limit and accident location. Thus dynamic board displays speed limit, accident location and time taken for ambulance to cross the particular dynamic board if it falls in the ambulance route. The idea is represented in Fig 2.

E. Detecting uneven roads

The normalized speed value is used to predict presence of uneven roads, the real time inductive loop values are compared along with the predicted normalized speed value based on the available inductive loop values as explained earlier. If the variation of the speed is more than 30% it means anomalies are predicted in a particular locality. The anomalies could be (4)

- Accident
- Occurrence of an event in or Around the locality
- Presence of uneven roads

For avoiding the other 2 possibilities if the variations are existing for more than a day are checked. On the assumption that event doesn't occur for more than a day and also compare real time people movement and traffic flow for the particular square id with the predicted people movements and traffic flow. If the variation between these elements are found to be very high (threshold value for the variation is set to be 30 percentage) then it implies that people movement is increased in the locality and results in reduction of speed and if the variation is less, it shows the people movement in the locality is the same as earlier predicted and hence this leads to possibility of uneven road. Thus the roadway authorities can be informed on the same by using the details provided under administrative region.

F. Rerouting

Rerouting is simply suggesting the backup path when the primary path fails, here occurrence of accident or unexpected congestion in a locality leads to failure of the primary path. Occurrence of accident is detected with the help of GPS and GSM inside a car. The accident detection model was explained earlier in system overview. Unexpected congestion is determined with the help of CDR details which predicts number of people. Any huge increase in number of people is considered as congestion and a rerouted path is suggested to the user. For implementation the same sample map created earlier for the time estimation model is used.

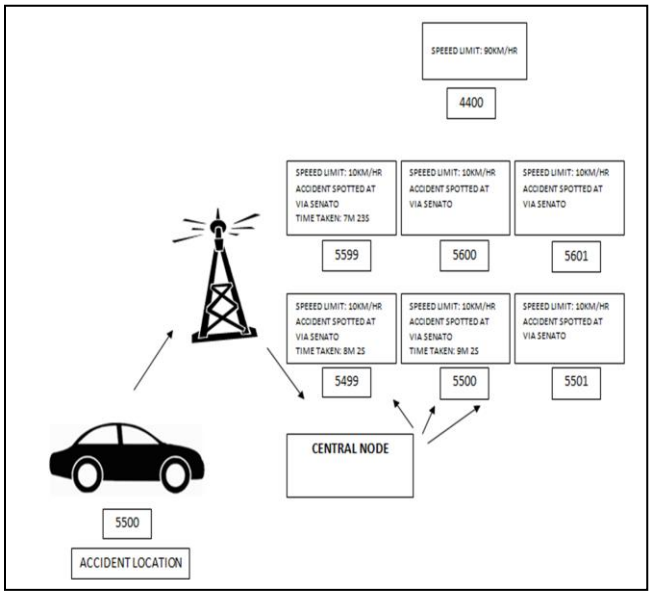


Fig. 2. **Traffic congestion avoidance during accident.** When central node is intimated about the accident directly from the vehicles, it finds the square id in which accident occurs with the help of latitude and longitude provided to them by the GPS inside car. Central node also picks the nearby square id and sends the message of new speed limit, accident location and time taken for the ambulance to cross the particular dynamic board (if the square id falls in the ambulance route) as shown in square id 5500 and sq id 4400 is far away from the accident location and hence speed limit doesn't varies

The rerouting concept is achieved by following steps.

a) Time taken is assigned as weights between nodes, in the form of distance by time.

b) If accident or any unexpected congestion is faced, central node allocates speed limit as 10km/hr which is updated in navigation tool between that particular nodes.

c) This implies time taken between the particular nodes will increase and subsequently the particular path is avoided by the navigation tool and alternative path is suggested.

VIII. Results and Discussion

In data analysis, as mentioned earlier, the available data is processed and stored in an hourly basis, separately for every square id. The parameters are traffic flow, approximate number of people, normalized speed and speed limit. Fig 3A shown here is analysis of output for square id 2266. It can be inferred that, when people's movement and traffic increases, the speed limit drops down. During midnight, the speed limit stays high and it toggles from 70 to 45 during peak hours indicating the betterment and more effectiveness of the dynamic speed board. A plot to illustrate variations in speed limit based on traffic flow and people movement is shown in Figure 3B. Estimating effective time between 2 destinations requires a sample map and shortest path between users chosen directions for our estimation. A Google map [21] is used for comparison. Comparison between Google's predicted time and the model prediction is shown in Fig 4. But time predicted by the model is more efficient only when a good number of Inductive loop sensors are used rather than private insurance company details since the details provided by them are based on fewer number of vehicles. The requirement of GPS installed in vehicles being mandatory on these days, dependency on private insurance company details will also increase.

Traffic congestion model is similar to controlling the mobility of vehicles from the center. Several papers have been published under this domain and few have been implemented too and this promises to be future in traffic avoidance. Here only a single case either accident or occurrence of big event which leads to unexpected traffic snarl up is considered. A solution is provided for the same by controlling the flow of vehicles with the help of dynamic board and rerouting vehicles through navigation tools. With the help of LabView [19], an automated environment is simulated. When accident is spotted, alert is sent to the central node and it sends details about new speed limit, accident location information and time taken to cross the dynamic board if it lies in ambulance route to all nearby square ids. The representation of city in square id has eased the difficulties in spotting the nearby square id and hence the dynamic board. The screenshot of traffic congestion avoidance model is shown in Fig 5. This helps in controlling the flow of vehicles and it also helps the ambulance to reach the destination (for eg. accident spot) in time.

Only one way route is considered because this would let victim to get first aid access more quickly. Adding to that, rerouted path helps in reducing the vehicles queuing and ultimately reduces the traffic snarl up around accident location. The simulation of sample scenario using MATLAB [18] is givens in Fig 6.

IX. Conclusion

An intellect central node along with dynamic speed board helped in addressing problems like proper management of vehicle movements, effective system for controlling the mobility of vehicles and understanding uneven roads and traffic snarl-up. Our proposed dynamic board set up at various locations in a city limits the vehicular speed and ensures control over vehicle's mobility. By considering all the real time parameters like traffic, weather or other anomalies, the efficient travel time between the destinations is estimated, which will help the user or the traveler to schedule plans accordingly. Vehicles are efficiently managed during anomalies like accidents by controlling the flow of vehicles, by reducing the speed limit and rerouting the local travelers by displaying the location of the accident in the dynamic board and by showing the rerouted path for other travelers using the navigation tool. To reduce the time taken for ambulance to reach the required destination the time taken for the ambulance to cross the dynamic boards is displayed in our approach. By displaying the time it is easy for the travelers to make way for the ambulance when necessary. Presence of uneven roads is always a hindrance to the travelers and halts the regular flow of the vehicles. With the help of the available sensors and M2M communication a cost efficient model is designed in this paper to predict the presence of uneven roads and report to the roadway authorities.

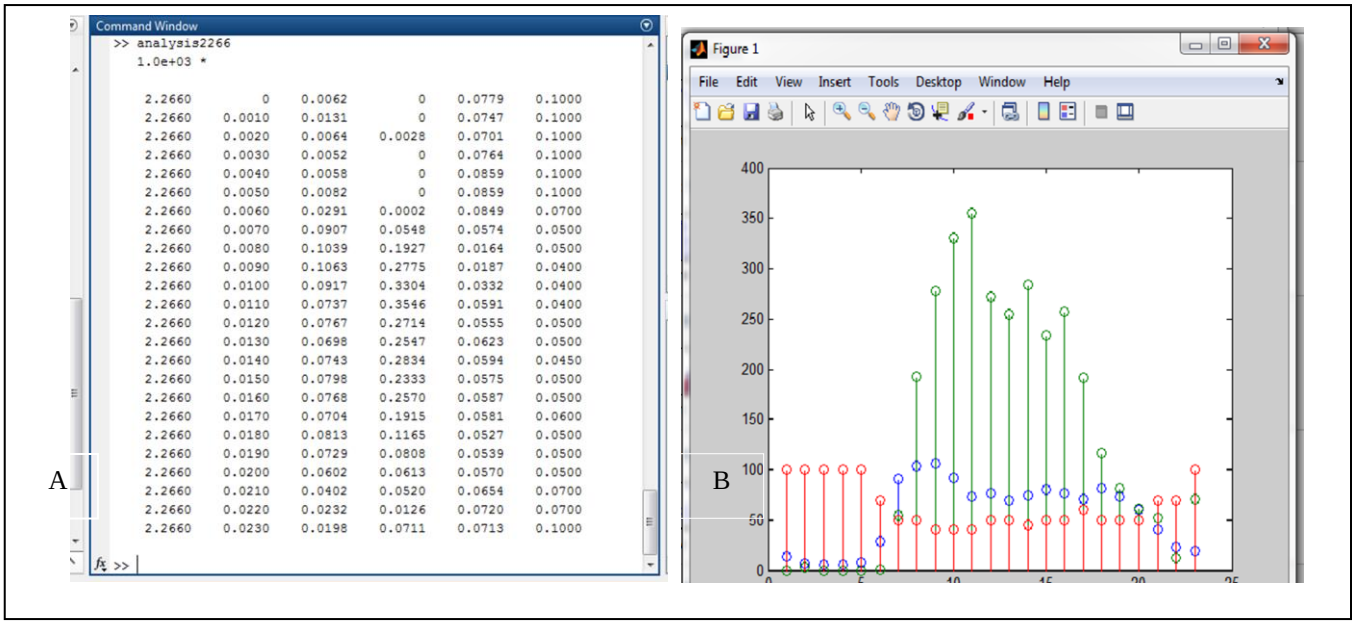


Fig. 3. **Data analysis for square id 2266 and Comparison of speed limit with traffic flow and number of people**. Figure 3A shows the Matlab analysis output for square id 2266, first column represents square id, sec column indicates time (0 represents 00:00 to 01:00 and so on),third column represents traffic flow (its average number of vehicles in the particular square id, calculated based on relativity), fourth column is the approximate estimation of number of people based on CDR details(estimated based on relativity), fifth column is the normalized speed (median value) in that particular square id and last column is the speed limit estimated based on traffic and number of people. In Figure 3B red line indicates speed limit, blue indicates traffic flow and green for number of people. The speed limit is at maximum value 100 if traffic flow and people movement values are small and the value dropping down if number of vehicles and people movement increases

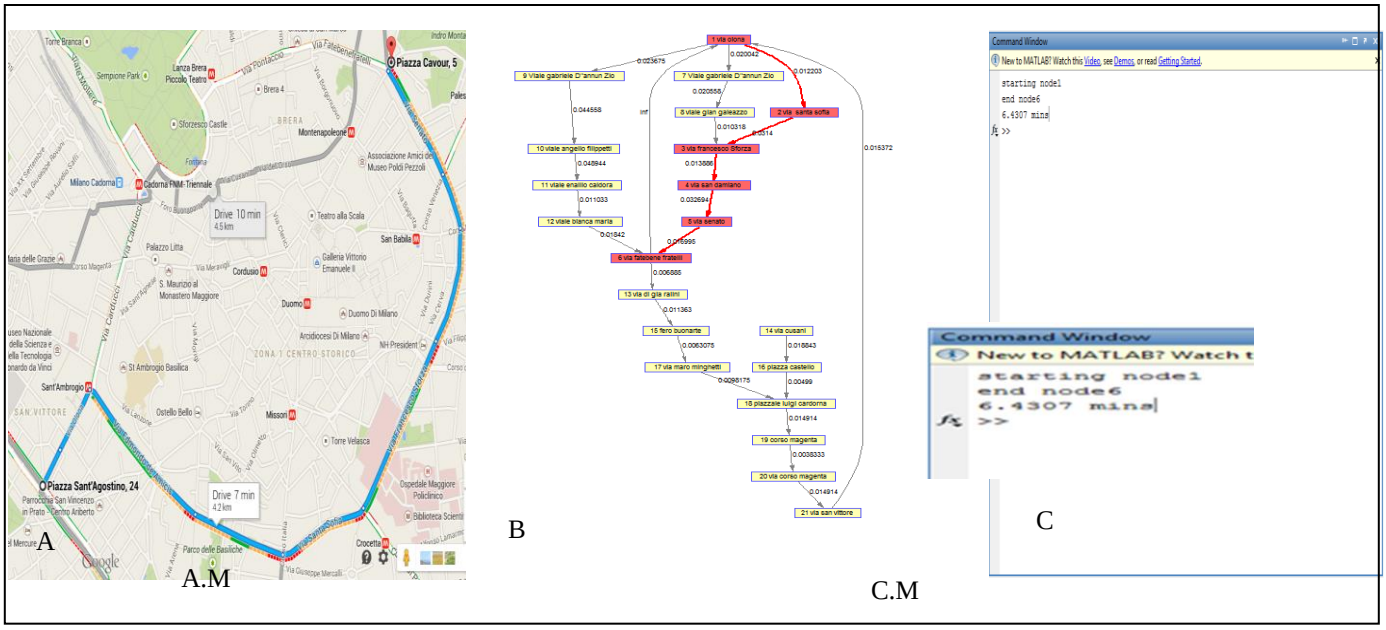


Fig. 4. **Comparison between Google and the time estimation is shown.** Figure 4.A shows google map shortest path and its estimated time taken, magnified image is shown in 4.A.M and Figure 4.B and 4.C shows the shortest path and time taken respectively, 4.C.M is the magnified version of 4.C. The Google predicted time is 7mins between via olona and via Fatenbenefratelli and the model predicted time is 6.4307mins. Though the model is in early stages, GPS installed in car getting mandatory in future it improvises the accuracy of the time estimation

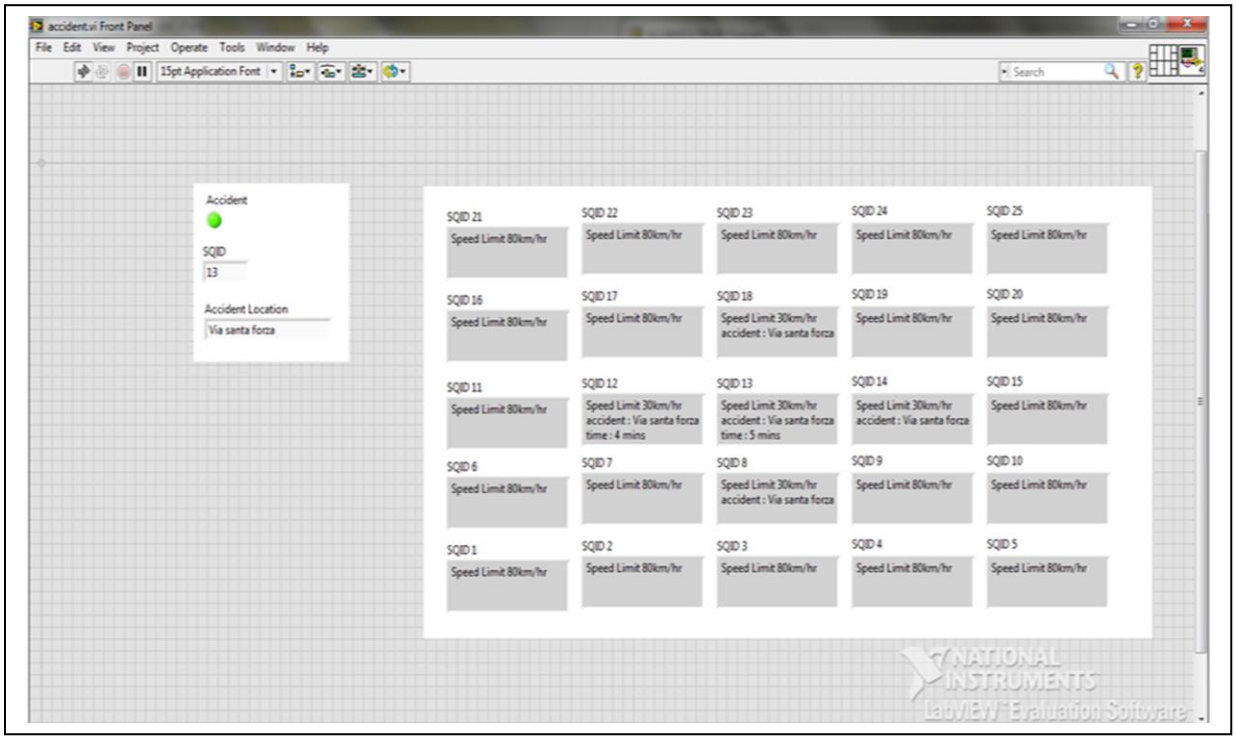


Fig. 5. **Traffic congestion avoidance model during accident.** Square id of 5x5 matrix is considered to depict the traffic avoidance congestion model during accident. The speed limit drops from 80 km/hr to 30 km/hr in accident spotted square id (13) and its neighbor square id (8, 12, 14 and 18), which controls the flow of the vehicle and reduces traffic snarl up. Adding to that if square id is located in the ambulance route then time taken for ambulance to cross the particular square id is displayed (square id 12 and 13). This would help ambulance to reach accident location on time

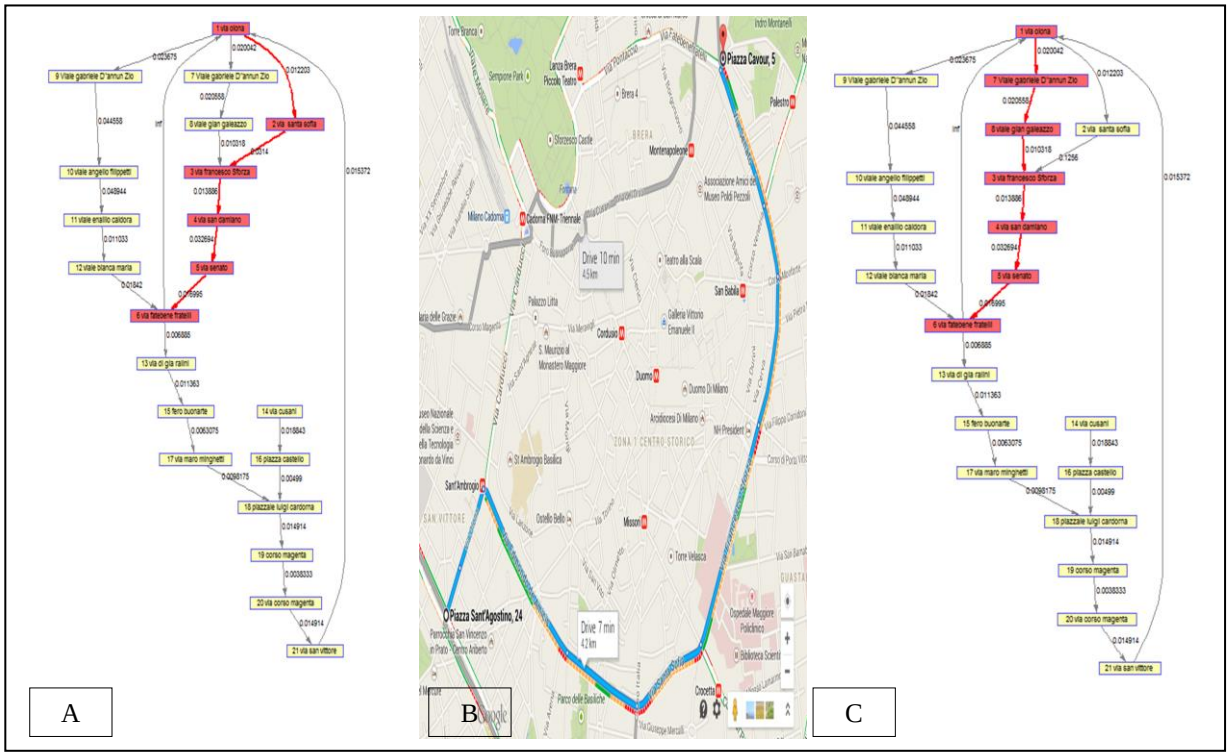


Fig. 6. **Rerouted path during accident and when people movement increases drastically from normal in a particular square id.** 6.A is the google hardest path between via Olona and via Fatenbenefratelli, 6.B is the shortest path between via Olona and via Fatenbenefratelli and the path taken is the same in both images proving the validity of the shortest path model and 6.C is the rerouted path when accident is spotted at via Santa Sofia, it reroutes the path via Viale **Gabriele Dannum Zio and Viale Gian galeazzo to reach via Francesco Sforza instead of Santa Sofia**

REFERENCES

[1] Understanding the Networked Society: New Logics for an Age of Empowerment, Ericsson White Paper, UEN 284 23-3242, Feb, 2015 http://www.slideshare.net/Ericsson/wp-understandingthenetworked society

[2] Fred Mannering, Walter Kilareski and Scott Washburn, Principles of Highway Engineering and Traffic Analysis, Wiley-India, 3rd ed, 2005.

[3] Kyeong-Pyo Kang, Gang-Len Chang, and Nan Zou. "Optimal Dynamic Speed-Limit Control for Highway Work Zone Operations," Transportation Research Record: Journal of the Transportation Research Board, No. 187, TRB, National Research Council, Washington, D.C., 2004, pp. 77-84

[4] Lu Xutao, "Design of transport vehicles remote monitoring system," 2nd International conference on Education Technology and Computer (ICETC), Vol 2, pp. 310-313, 2010.

[5] M. Jason S. Claire, M. Stephen, and T. Denis, "3D laser imaging for surface roughness analysis". International Journal of Rock Mechanics and Mining Sciences, Vol. 58, pp. 111-117, 2012

[6] Q. Li, M. Yao, X. Yao, and B. Xu, "A real-time 3D scanning system for pavement distortion inspection," Journal of Measurement Science and Technology, Vol. 21(1), pp. 15702-15709, 2010.

[7] Z. Hou, K. Wang, and W. Gong, "Experimentation of 3D pavement imaging through stereovision," Proceedings of International Conference on Transportation Engineering, pp. 376-381, 2007.

[8] X. Yu and E. Salari, "Pavement pothole detection and severity measurement using laser imaging," Proceedings of 2011 IEEE International Conference on Electro/Information Technology, pp. 1-5, 2011.

[9] K. Christian and B. Ioannis , "Pothole detection in asphalt pavement images." Journal of Advanced Engineering Informatics, Vol. 25 (3), pp. 507-515, 2011.

[10] B. Yu, and X. Yu, "Vibration-based system for pavement condition evaluation," Proceedings of the 9th International Conference on Applications of Advanced Technology in Transportation, pp. 183-189, 2006.

[11] In Jung Lee, "An accident detection system on highway through CCTV with calogero-moser system," 18Th Asia pacific conference on Communication (APCC), pp. 522-525, 2012.

[12] C. Thompson. J. White. B. Dougherty, A. Albright, and D.C. Schmidt, "Using smart phones to detect car accident and provide situational awareness to emergence responders," 3rd international ICST conference in Mobile wireless Middleware, Operating system and applications, Mobilware 2010, LNICST 48, pp. 29-42, 2010..

[13] M.S. Amin, J. Jalil, and M.B.I. Reaz, "Accident detection and reporting system using GPS, GPRS, and GSM technology," 2012 International conference on Informatics, Electronics and Vision (ICIEV), pp. 640-643, 2012.

[14] J. Pan, M.A. Khan, I.S. Popa, K. Zeitouni and C. Borcea, "Proactive Vehicle Re-routing Strategies for Congestion Avoidance," 2012 IEEE 8th international conference on Distributed Computing in Sensor Systems (DCOSS) pp. 265-272, ,2012.

[15] Siwei Jiang, Jie Zhang, and Yew-Soon Ong, "A Pheromone-based Traffic Management Model for Vehicle Re-routing and Traffic Light Control," Proceedings of the 2014 International Conference on Autonomous Agents and Multi-agent Systems, pp. 1479-1480, 2014.

[16] Lili Huang and Matthew Barth, "A Novel Loglinear Model for Freeway Travel Time Prediction," Proceedings of the 11th International IEEE Conference on Intelligent Transportation Systems, pp. 210-215, 2008.

[17] http://www.d4d.orange.com/en/home

[18] http://in.mathworks.com/products/matlab/

[19] http://www.ni.com/labview/

[20] http://in.mathworks.com/matlabcentral/fileexchange/27785-distance-calculation -using-haversine-formula/content/haversine.m

[21] https://www.google.com/maps/d/viewer?mid=zNustSkyY4mQ.kbAGthcvLJu0&ie=UTF8&hq&hnear=Italy&source=embed&oe=UTF8&showlabs=1&msa=0

Developement of Bayesian Networks from Unified Modeling Language for Learner Modelling

ANOUAR TADLAOUI Mouenis
LIROSA
Faculty of Sciences, Abdelmalek Essaadi University
Tétouan, Morocco

AAMMOU Souhaib
LIROSA
Faculty of Sciences, Abdelmalek Essaadi University
Tétouan, Morocco

KHALDI Mohamed
LIROSA
Faculty of Sciences, Abdelmalek Essaadi University
Tétouan, Morocco

***Abstract*—First of all, and to clarify our purpose, it seems important to say that the work we are presenting here lie within the framework of learner modeling in an adaptive system understood as computational modeling of the learner .we must state also that Bayesian Networks are effective tools for learner modeling under uncertainty. They have been successfully used in many systems, with different objectives, from the assessment of knowledge of the learner to the recognition of the plan followed in problem solving. The main objective of this paper is to develop a Bayesian networks for modeling the learner from the use case diagram of the Unified Modeling Language. To achieve this objective it is necessary first to ask the Why and how we can represent a Learner model using Bayesian networks? How can we go from a dynamic representation of the learner model using UML to a probabilistic representation with Bayesian networks? Is this approach considered experimentally justified? First, we will return to the definitions of the main relationships in the diagram use cases and Bayesian networks, and then we will focus on the development rules on which we have based our work. We then demonstrate how to develop a Bayesian network based on these rules. Finally we will present the formal structure for this consideration. The prototypes and diagrams presented in this work are arguments in favor of our objective. And the network obtained also promotes reusing the learner modeling through similar systems.**

Keywords—Learner Modeling; Bayesian networks; Cognitive diagnosis; Uncertainty

I. INTRODUCTION

The problem of this paper can be summarized as follows: Why and how can we represent a Learner model using Bayesian networks? How can we go from a dynamic representation of the Unified modeling language (UML) model to a probabilistic representation with Bayesian networks? Is this consideration experimentally justified?

The learner model is a data structure that represents the state of knowledge of a learner in a given field. This model identifies the learner's current level of understanding of the domain knowledge. It includes data on individual variables of a learner that allow updating of the learner profiles from information obtained during the interactions.

All existing approaches to model the learner are based generally on using the Unified Modeling Language [1], that quickly became a standard for the analysis and design in software development. It provides a schematic approach to describing the needs of the user, which begins with the use cases diagrams, and leads to a more formal specification, using stereotyped classes in the analysis model. The components of this modeling language form the basis of an architectural view in the system while providing the foundation for the design, implementation and validation and verification.

We have attempted in previous works, to model the learner using Bayesian networks [2] and multi networks [3] as a formalism to manage uncertainty in the management of learner model. In this paper, we will try to offer a combination of these two approaches, starting with specifying the transformation rules on which we have based our work. We will then demonstrate how to transform the use case diagrams into a Bayesian network based on these rules. Finally we will present the formal structure for this consideration.

II. USE CASE DIAGRAM, UML POINT OF VIEW

A. Definition

Use cases describe the form of actions, reactions and the behavior of a system from a user perspective. They allow defining the limits of the system and the relationship between the system and the environment.

Use cases are filling a lack of raw object methods, such as [4] and [5], which did not offer techniques, for the identification of needs. In this sense, the use cases associated with technical objects allow a comprehensive approach to the entire life cycle, from the specification to implementation.

A use case is a specific way of using a system. It is the image of system functionality, triggered in response to the stimulation of an external actor.

B. Use case diagram's Relationships

Use case diagrams represent actors and relationships between actions and actors. We will define in this section the main relationships in the use case diagram that we will use in our work, they are three types of relationships between actors and use cases:

1) Generalization relationship

A case A is a generalization of a case B if B is a particular case of A. For example, the consultation of an account via the Internet is a particular case of the consultation. This relationship of generalization / specialization is present in most

of the UML diagrams and results in the concept of inheritance in object-oriented languages.

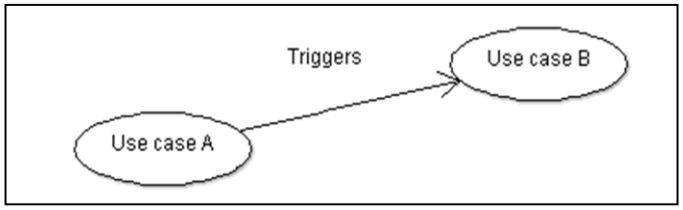


Fig. 1. Generalization relationship in a use case diagram

"Figure. 1", shows that use case A initiates a use case B; this action is represented by an arrow from the initiator of the action to the triggered action.

2) Inclusion relationship

A case A includes a case B if the behavior described by the case A includes the conduct of the case B: Where A depends on B. When A is applied, B also must be applied as a part of A.

This dependence is symbolized by the "include" stereotype. For example, accessing information from a bank account necessarily includes an authentication phase with a username and password.

The inclusions essentially allow factorizing a part of the description of a use case that would be common to other use cases like "Figure.2" shows. The inclusions are also used to decompose a complex case into simpler sub-cases.

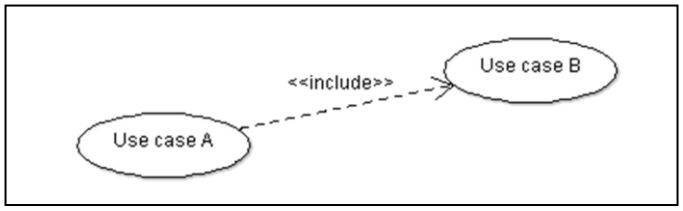


Fig. 2. Inclusion relationship in a use case diagram

3) Extension relationship

It is said that use case A extends a use case B when the use case A can be called during the execution of the use case B. Running B may possibly lead to the execution of A: unlike the inclusion, the extension is optional. This dependence is symbolized by the stereotype "extend"

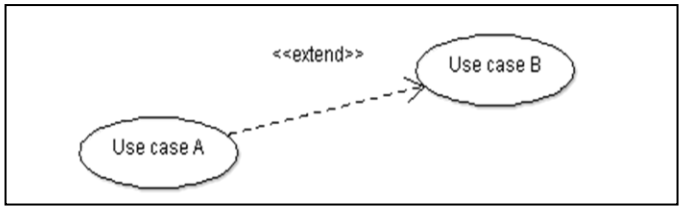


Fig. 3. Extension relationship in a use case diagram

III. Learner Model

A. Definition

A learner model allows keeping the learner information, for example his level of knowledge on a given topic (performance), his frequent mistakes/misunderstandings, psychological characteristics, etc.

A learner model can be defined as a set of structured information about the learning process, and this structure contains values on the characteristics of the learner. [6] It provides the necessary data to the other modules to achieve the adaptation of teaching to the learner. [7]

Many studies emphasize the uncertainty of the information contained in the student model and the importance of the intention behind the creation of this model. Thus, a student model represents the belief system about learners' beliefs (the system's beliefs about the learner's beliefs) accumulated during the diagnostic process.

B. Typologies of Learner model

The learner model is a data structure in the computer sense that characterizes for the learning environment, the state of a subset of the learner's knowledge from the system point of view.

It will be defined by the difference between the learner knowledge and target knowledge, issue of learning, as represented in the system. The approach to represent this difference leads to distinguish two major classes of models:

- The models of partial or overlay expertise [8], in which the knowledge of the learner is only a subset of the target knowledge. The idea behind this type of model is that the learner present deficiencies or the poorly insured knowledge, or somehow weaknesses, it is identified to allow it to grow. The aim of the learning system is then to complete the knowledge of the learner in order to acquire all the knowledge outlined in the model.
- Differential models [9], which incorporate "false knowledge", corresponding to perturbations of the expert knowledge or erroneous preconceptions. In fact, studies show that many errors are not due to erratic behavior of learners, but the correct application of false procedures. To develop a model of learners' knowledge, one must take into account these types of systematic errors, that researchers will be designated by the term "bug" (bug).

While a partial model invites expertise in teaching strategies centered on the fact to fill the gaps of the learner, the incremental models will lead to strategies based on remediation.

IV. Bayesian Networks

A. Definition

Knowledge representation and reasoning from these representations has created many models. Probabilistic graphical models, specifically Bayesian networks initiated by [10] in the 1980s, have proven to be useful tools for representing uncertain knowledge and reasoning from incomplete information.

A Bayesian network $B = (G, N)$ is defined by

- $G = (X, E)$, acyclic directed graph with vertices associated with a set of random variables $X = (X, \dots, Xn)$;

- $N = \{P(Xi \mid Pa(Xi))\}$ All the probabilities of each node Xi conditionally to the state of its parents $Pa(Xi)$ in G.

Thus, the graphical part of the Bayesian network indicates the dependencies (or independence) between variables and provides a visual tool for knowledge representation, more easily comprehensible by its users. In addition, the use of probability allows taking into account uncertainty in quantifying the dependencies between variables. Both properties have been the cause of the first names of Bayesian networks, "probabilistic expert systems", where the graph was compared to the set of rules of conventional expert system, and the conditional probabilities presented as a quantification uncertainty about the rules.

[11] Also have shown that Bayesian networks allow representing compactly the joint probability distribution over the set of variables:

$$P(X1, X2, \cdots, Xn) = \prod_{i=1}^{n} P(Xi \mid Pa(Xi))$$

This decomposition of a global function as a local product terms depending only of the node and its parents in the graph, is a fundamental property of Bayesian networks. It is the basis of the first work on the development of inference algorithms which calculate the probability of any variable of the model from the same partial observation of other variables. This has been proven NP-complete, but resulted in different algorithms that can be treated as information propagation methods in a graph. These methods obviously use the concept of conditional probability, i.e. what is the probability of Xi knowing that I have observed Xj, but also the Bayes theorem, that calculates, conversely, the probability of Xj knowing Xi, when $P(Xi \mid Xj)$ is known.

B. Construction of a Bayesian network

As we have seen in the definition, the complete specification of a Bayesian network requires specifying a share structure (directed acyclic graph that underlies) and other parameters (probability tables). To do this, two approaches are possible and can be combined: the collection of expertise and machine learning, which is one of the attractions of Bayesian networks.

In the case of collection of expertise, the definition of the network structure begins with the identification of possible nodes and the distinction between (unobservable) informational variables (inputs) or hypothetical. The existence of an arc can be analyzed in terms of influence of one variable on another, but its orientation is more difficult. Traditionally, an arc is directed from A to B if A is a cause of B, but we will see that this interpretation is not as simple in the case of the learner modeling. The parameters are in turn attached in an approximate manner by using frequentists or qualitative information.

Since Bayesian network is a probability distribution, we can use maximum likelihood as statistical learning parameters criterion. The result is as a Bayesian network whose structure is fixed and E which is a comprehensive basis of example, the maximum likelihood is achieved if the parameters of the Bayesian network are equal to the frequencies of the same features observed in E. statistical learning structure requires for its development test to determine whether or not the random variables are conditionally independent [12].

V. Development of Bayesian Network from a Use Case Diagramme

A. The choice of Bayesian networks

As we previously presented, the diagrams of use cases is a top view of system features, it allows us to present all user actions (learner in our case). These actions may require elements of uncertainty, this uncertainty will clearly present when poised to collaboration diagrams. Representing this uncertainty becomes very important when there are a large number of interdependent and potentially conflicting requirements that overwhelm the capacity of spontaneous human spirit.

Bayesian network models explicitly the uncertainty between the requirements represented by use case and collaboration diagrams elements. During the presentation of the functional evidence such as the importance of a particular learner, a quantitative assessment can be performed to the way we strongly believe the requirement is indicated. We therefore see the ability to transform the use case diagrams of Bayesian networks as a significant potential lead in the modeling of the learner.

We believe that Bayesian networks will provide a solution that will allow us to understand and measure a dynamic way all the actions of the learner in a learning situation. Networks obtained, we will give a capacity to monitor and represent at real time, all the actions of the learner, the rationale for these choices, and identification of each of the paths that will be followed during a learning situation.

B. Bayesian network development's rules

1) Generalization relationship

A generalized use case diagram contains a common functionality that is available for all the specialized use cases. The transformation of the generalization relationship to the nodes of a Bayesian network is simple:

Consider "Figure.4" in the use case A is a generalization of the use case A1 and the use case A2, we represent the functional requirement A1 and A2 being a descendant of the functional requirement of A.

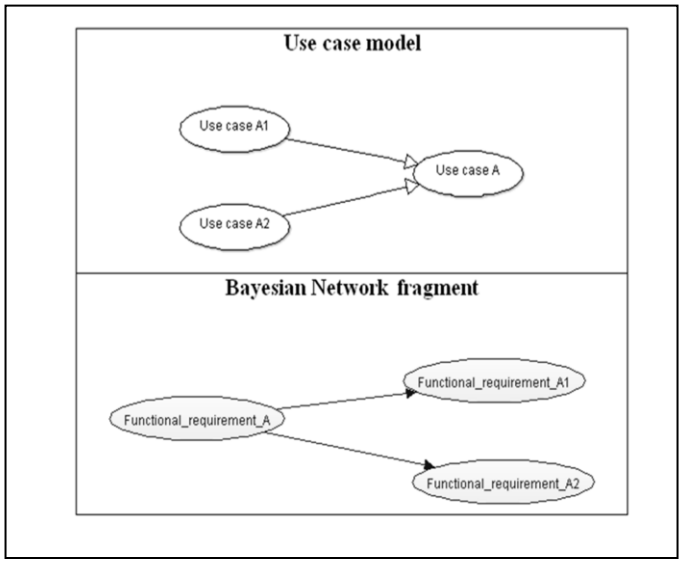


Fig. 4. Bayesian fragment developed from the generalization relationship

This results in a Bayesian network with a similar structure. The direction of the arc flow from A to A1 and A2 reflecting a top-down decomposition. The information represented in the arrows of the use case will be included in the functional requirements. This indicates that it is more likely to encounter the general case of the specific functional requirement. Thus:

$$P(A) = prior$$

$$P(A1 \mid A) = P(A \mid A1)P(A1) / P(A)$$

$$P(A2 \mid A) = P(A \mid A21)P(A2) / P(A)$$

2) Inclusion relationship

The inclusion relation in a use case diagram models the situation in which a use case is composed of a desired number of use cases. For inclusion, the high level of use cases cannot run without the implementation of sub use cases.

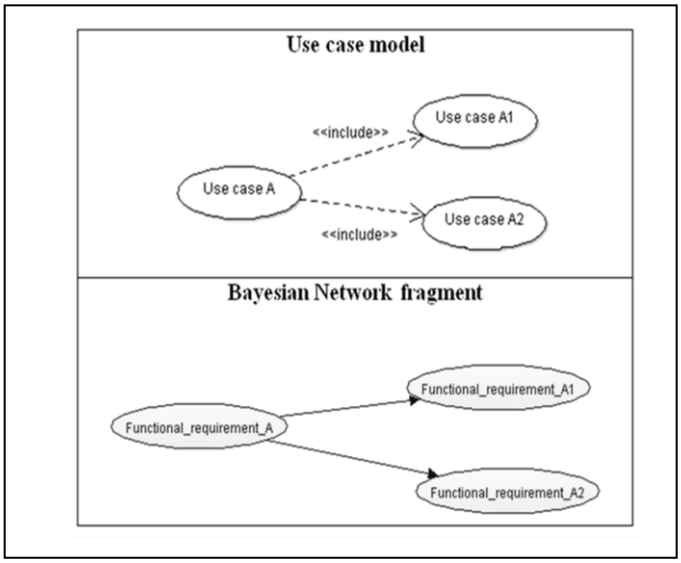


Fig. 5. Bayesian fragment developed from the inclusion relationship

To see how this can be translated to fragments of the Bayesian network, consider "Figure.5" use case A is linked to use cases A1 and A2 by an inclusion relation

This results in a Bayesian network with a similar structure as the generalization relationship. The direction of the arc flow from A to A1 and A2 reflecting a composition from bottom to top in the information represented in the arrows of the use case will be included in the functional requirements. This indicates that it is more likely to encounter the general case of the specific functional requirement. Thus:

$$P(A) = prior$$

$$P(A1 \mid A) = P(A \mid A1)P(A1) / P(A)$$

$$P(A2 \mid A) = P(A \mid A21)P(A2) / P(A)$$

3) Extension relationship

The extension relationship in a use case diagram represents a particular use case branched additional behavior given the satisfaction of certain conditions. In case of extension, the first use case does not need any more use case to run. The second use case is an exceptional behavior if the conditions are fulfilled.

Consider the general case schematized on "Figure 6". A use case is extended by the case of A1 use. This models the situation in which an additional criterion triggers the case of using A1 after executing use case A.

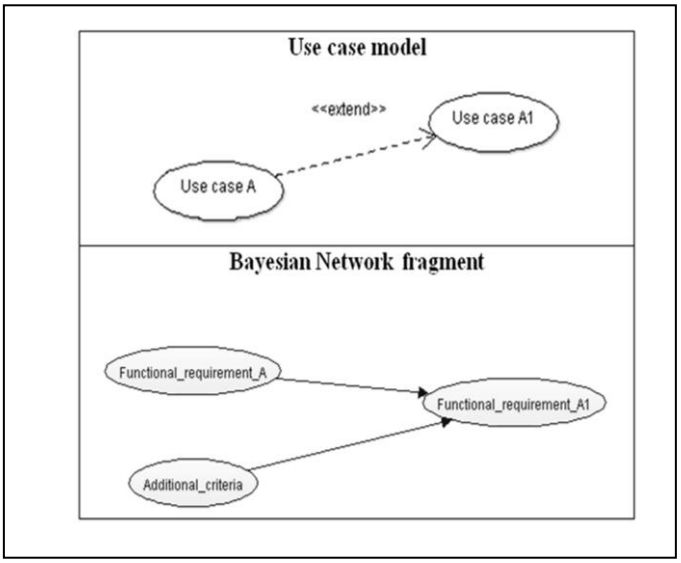


Fig. 6. Bayesian fragment developed from the extension relationship

The additional criterion is described in the flow of events from textual description. This situation is modeled as functional requirement A1 implied by the functional requirement A. The additional criterion is modeled as another functional requirement node. The direction of the implication is the additional criterion (AC) to the functional requirement A1. Thus:

$$P(A1 \mid A, AC) = \frac{P(A|A1, AC)P(A1|AC)}{P(A|AC)}$$

VI. Learner Modelling Using Bayesian Networks

Our work lies in the framework of learner modeling in an adaptive educational system, to illustrate the ideas discussed in the previous sections; we will focus our work on the actions of a learner in an adaptive system. We defined the "Table 1" several actions of a learner in a learning situation.

TABLE I. Learner Action in an Adaptive System

Learner's actions
• Post question in the forum. • Follow courses. • Take the pretest.

A. Learner use case diagram

Considering "Figure.7" a main actor is identified, named the learner. The figure shows the generalization relationships between use cases and the learner, and generalization relationships, inclusion and extension between use cases.

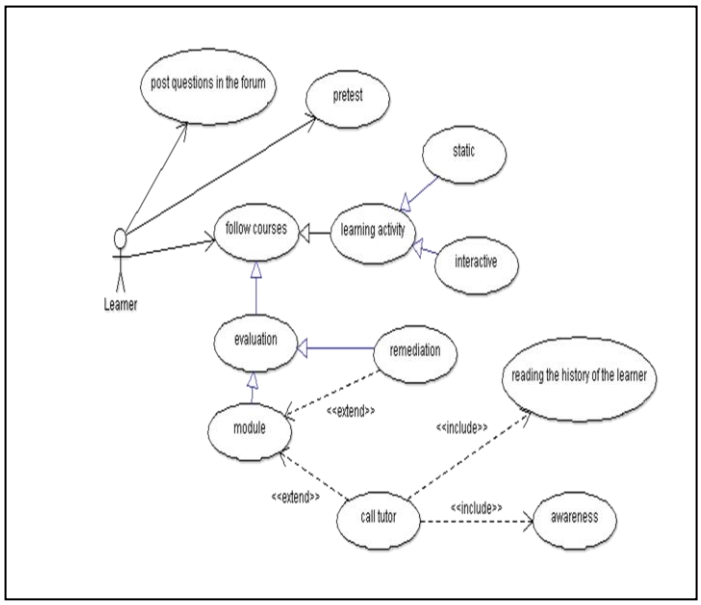


Fig. 7. Use case diagram of a learner model's actions in an adaptive system

In particular, the functional requirement "follow courses" is represented with a generalization relationship between the functional requirements "learning activity" and "evaluation". The functional requirements "post in the forum issues" and "pretest" are represented with a generalization relationship with actor "Learner". There are also extensions relationships in the functional requirement "module" and its relationship to functional requirements "remediation" and "call tutor." Inclusion relations are presented in the representation of the relationships between functional requirements "call tutor", "reading the history of the learner" and "system awareness."

B. Bayesian network obtained

Once the use case diagrams have been created, it is easy to create the structure of the Bayesian network using the rules described in the previous sections. "Figure.8" represents the Bayesian network representation of the main actions of the learner in a Learning situation, constructed from the use case diagram shown in "Figure.7". Note how the conditional independence was directly modeled by applying the rules as shown.

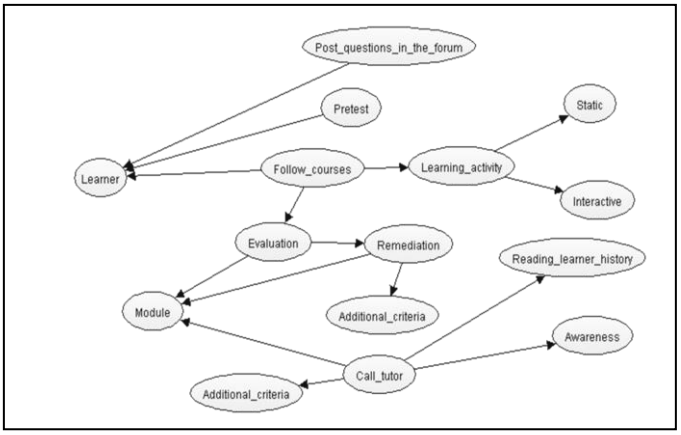


Fig. 8. Bayesian network developed from the learner model's use case diagram

VII. Conclusion and Perspectives

We have shown in this work, how we can develop with well defined rules; Bayesian networks from use case diagrams of Unified Modeling Language, the development of the Bayesian network in future work could be done manually using the rules we provides in previous section of our work, or in using a software that can provide this transformation automatically, we couldn't find during our research a software which grantee this transformation, but we are working in development of a tools which allow this transformation automatically.

This work is a major step in our research in modeling the learner in an adaptive educational system, the transition from the use case diagrams towards Bayesian networks; give us the opportunity to reach our goal to use the Bayesian networks as a formalism to manage uncertainty in the modeling of the learner.

We see two main directions in which we can continue this work. On the one hand by applying our ideas to more advanced conceptual models as used so far. And on the other hand is transforming the Bayesian networks obtained a machine readable language, or one using probabilistic ontology as we proposed in previous work [13], or perform a combination of Bayesian networks with ontologies.

References

[1] Booch, G., Rumbaugh, J. and Jacobson, I. The Unified Modeling Language User Guide. Addison-Wesley Publishing Company, Reading, MA, 1999.

[2] M. Anouar Tadlaoui, M. Khaldi, S. Aammou (2014) Towards a Learning model based on Bayesian Networks,EDULEARN14 Proceedings, pp. 3185-3193.

[3] M. Anouar Tadlaoui, M. Khaldi, S. Aammou (2014) Bayesian Networks for Learner Modeling, International Journal of Basic Sciences and Applied Computing 1 (1), 5-9.

[4] James Rumbaugh, Michael Blaha, William Premerlani, Frederick Eddy, William Lorensen (1990). Object-Oriented Modeling and Design. Prentice Hall

[5] Jacobson, Ivar; Grady Booch; James Rumbaugh (1998). The Unified Software Development Process. Addison Wesley Longman.

[6] L.Zaitseva, C.Boule. Learning systems in professional training. Workshop "Industry meets research" within the conference Interactive Computer Aided Learning ICL 2005 Villach, Austria 28 – 30 September 2005

[7] Beck, J., Stern, M., & Haugsjaa, E. (1996). Applications of AI in education. Crossroads, 3(1), 11-15.

[8] Brian Carr, Ira P. Goldstein. Overlays: a theory of modelling for computer aided instruction. AI Memo 406. 1977.

[9] Wenger, E. (1987). Artificial Intelligence and Tutoring Systems. Morgan Kaufmann, Los Altos, CA 94022.

[10] [PEARL, J., Probabilistic Reasoning in Intelligent System, Morgan Kaufmann. 1988.

[11] Geiger D., Heckermann D. (1996). Knowledge representation and inference in similarity networks and Bayesian multinets, Artificial Intelligence, volume 82 (1–2), pp. 45–74.

[12] BECKER, A., NAÏM, P., Les réseaux Bayésiens, modèles graphiques de connaissances, Eyrolles, 1999.

[13] M. Anouar Tadlaoui, M. Khaldi, S. Aammou (2014) Towards Probabilistic Ontology based on Bayesian Networks, International Journal of Software and Web Sciences 1 (10), 102-106.

High Accuracy Arabic Handwritten Characters Recognition Using Error Back Propagation Artificial Neural Networks

Assist. Prof. Majida Ali Abed
College of Computers Sciences & Mathematics,
University of Tikrit, Tikrit, Iraq

Assist. Prof. Dr. Hamid Ali Abed Alasad
Computers Sciences Department, Education for Pure Science College, University of Basra, Basra, Iraq

***Abstract*—This manuscript considers a new architecture to handwritten characters recognition based on simulation of the behavior of one type of artificial neural network, called the Error Back Propagation Artificial Neural Network (EBPANN). We present an overview of our neural network to be optimized and tested on 12 offline isolated Arabic handwritten characters (ا,ت,ج,ذ ،ش،ط،غ،ف،ك،م،و،ي) because the similarity of some Arabic characters and the location of the points in the character. Accuracy of 93.61% is achieved using EBPANN which is the highest accuracy achieved during Offline Handwritten Arabic Character Recognition. It is noted that the EBPANN in general generates an optimized comparison between the input samples and database samples which improves the final recognition rate. Experimental results show that the EBPANN is convergent and more accurate in solutions that minimize the error recognition rate.**

***Keywords*—Character Recognition; Neural Network; Classification; Error Back Propagation Artificial Neural Network**

I. INTRODUCTION

Many types of Artificial Neural Network have been proposed over the years. In fact, because Artificial Neural Network (ANNs) are studied by many sciences such as Computer Scientists, Electronic Engineers, Biologists and Psychologists, and has many different names such as Artificial Neural Networks (ANNs), Connectionism or Connectionist Models, Multi-layer Perceptron's (MLPs) and Parallel Distributed Processing (PDP). An ANNs defines a mathematical model for the simulation of a network of biological nervous systems and one of the most active learning methods to approximate real-valued, discrete-valued, and vector valued functions. It has been widely and very good used in pattern recognition problems, such as handwritten characters, recognizing spoken words, and face recognition [1]. The Error Back Propagation Artificial Neural Network(EBPANN) is a very common model in artificial neural networks and it does not have feedback connections, but errors propagate backward from the output layer during training. Training a network by Error Back Propagation Artificial Neural Network (EBPANN) includes three steps [2, 3];

- Feed forward of the input training pattern.
- Backward propagation of the associated error.
- Modification of weights.

To modification of connection weights between the pairs of layers, the errors in the output determine measures of hidden layer output errors which are used. The process of modifying the weights between the layers and calculate again the output continues until a stopping criterion is satisfied, such as the total error is reduced and reached to the value of given threshold. After training is completed, the network can be used to find outputs for new inputs [4].Pattern recognition is important application of artificial neural network which can be applied by using a feed-forward artificial neural network Figure (1) explain this operation that has been trained accordingly.

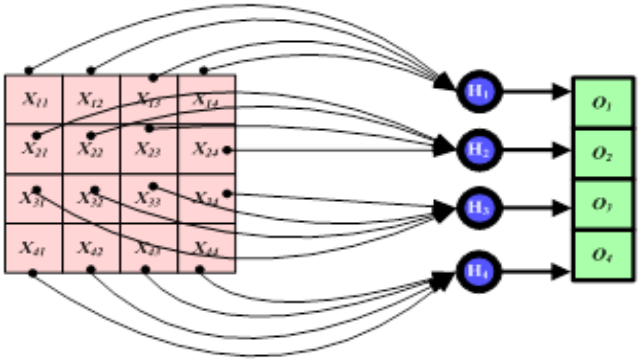


Fig. 1. Feed-forward artificial neural network for pattern recognition

Aim of this manuscript to implement the Error Back Propagation Artificial Neural Network (EBPANN) algorithm to optimize and Recognition the offline isolated Arabic handwritten characters. A brief review of neural network and Error Back Propagation Artificial Neural Network algorithm for training neural networks are presented in section 2. We present our proposed approaches, the pre-processing steps, the feature extraction technique and the classification in section 3. The knowledge base and experimental analysis are made in section 4. Results and discussion is presented in Section 5. Finally the conclusions are given in section 6.

II. ERROR BACK PROPAGATION ARTIFICIAL NEURAL NETWORK

Error Back Propagation Artificial Neural Network was first described by Paul Werbos in 1974, and further developed by David E. Rumelhart, Geoffrey E. Hinton and Ronald J. Williams in 1986 [5]. It was Common supervised learning technique; it's used for pattern recognition, function

approximation, prediction and Mapping tasks. The Error Back Propagation Artificial Neural Network algorithm has been widely used and a popular technique as a learning algorithm by examples in feed forward with a finite number of pattern pairs consisting of an input pattern and a corresponding (target) output pattern. Multiplayer neural networks each layer consists of one or more cells (neurons), each cell (neuron) in layer receives input from each cell (neuron) in the previous layer, and output to each cell (neuron) in the following layer. The Back propagation is applied to feed forward artificial neural network with one or more hidden layers this algorithm is different than others in the way in which the weights are calculated during the learning phase of the network. Zweiri, Y.H. et al [6] minimized the total squared error of the output based on Error Back Propagation Artificial Neural Network. The important points in the Error Back Propagation Artificial Neural Network are:

- Networks in layered structure (multi-layers network).
- The value of all weights are random
- Learns by examples and learning process is done through sequential mode and batch mode
- Error calculated on output layer and propagated back to hidden layers.
- Consists from two parts forward part(calculated output) reverse pass(actual output)
- The input and its corresponding output(target)are called a training Pair

Weight changes calculated using learning rate (η)

Error Back Propagation Artificial Neural Network has some problems which include network paralysis, local minima and slow convergence these problem occur when the network always change the weights, modify value of weights to very large value during the training operation. There are a number of solutions for these problems. The important one is very simple and that is to change the weights to different random numbers and try training again. Another solution is to add "momentum" to the weight Change. This means that a weight changes from iteration to iteration at any given time depends not just on the present error, but also on preceding changes. Further training of the Error Back Propagation Artificial Neural Network is continued till the desired classification performance is reached. The algorithm consist of two parts first part is forward propagation, second part is reverse propagation which explain below the two parts and Multilayer feed-forward networks are trained using Error back propagation Artificial Neural Network learning algorithm as shown in Figure (2) [7].

- Forward Propagation part:

Step 1: Compute δ_i^l in the output layer (oi= y_i^l).

$$\delta_i^l(c) = g'(h_i^l)\left[d_i^w(c) - y_i^l(c)\right] \quad (1)$$

Where h_i^l represents the net input to the ith cell in the Ith layer, and g' is the derivative of the activation function g.

Step 2: The error of a character is calculated

$$E(c) = (1/2)\sum\left[d_i^w(c) - \delta_i^l(c)\right]^2 \quad (2)$$

- Reverse Propagation part:

Step 3: Compute the deltas for the preceding layers by propagating the errors backwards;

$$\delta_i^l(c) = g'(h_i^l)\sum_j w_j^{l+1}\,\delta_i^{l-1}(c) \quad (3)$$

Where w_j^{l+1} the weights and the following equations define the change in the weights

$$w_{kj}^l(t+1) = w_{kj}^0(t) + \eta\,\delta_{pk}^0\,i_{pj}$$

$$w_j^{l+1}(t+1) = w_{ji}^l(t) + \eta\,\delta_{pj}^l\,x_i$$

Step 4: Update weights using equation (4);

$$\Delta w_{ji}^l = \eta\ \delta_i^l\,y_i^{l-1} \quad (4)$$

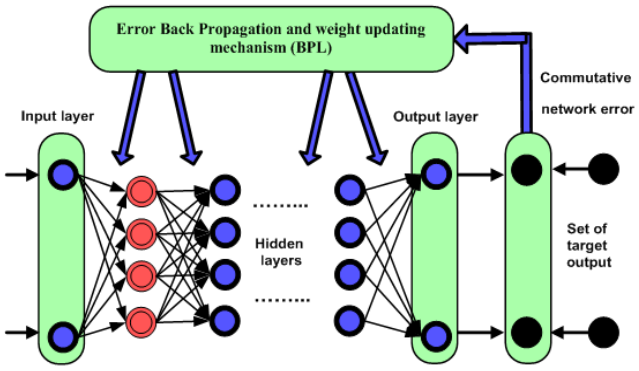


Fig. 2. Multilayer Feed-forward Neural Network.

III. Proposed Neural Network Structure Work

Our proposed neural network structure work offline Arabic handwritten characters recognition is implemented with the help of Error Back Propagation Artificial Neural Network (EBPANN). Handwritten character recognition system can be divided into four steps and shown in Figure (3):

1) *Character acquisition.*
2) *Pre-processing.*
3) *Feature extraction.*
4) *Classification and Recognition.*

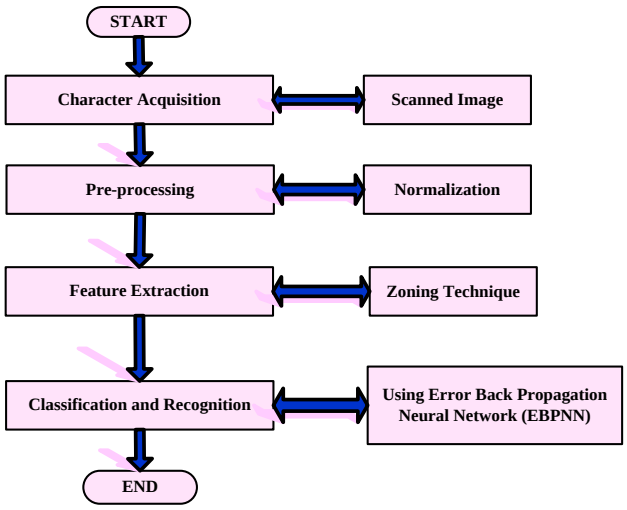


Fig. 3. Our neural network structure

1) Character acquisition

Acquisition of offline handwritten Arabic characters is done by using a scanner. In this investigation, some isolated Arabic handwritten characters been employed for image database in BMP type file. The Method of character acquisition in our neural network structure depend on the database used in the experimentation was preprocessed, size normalized images of scanned 252 image of offline handwritten Arabic characters.

2) Pre-processing

The process of normalization of characters it is necessary performed in pre-processing operation of our manuscript so that all characters can become of equal dimensions of the matrix, the size of a character varies but is typically around of 30 x30 pixels, since the input of Error Back Propagation Artificial Neural Network (EBPANN) is fixed size, it is necessary to normalize the size of the input characters and made the characters with same size, the normalization performed by using linear transformation to make the size of all characters with equal dimension in 18x18 array of binary pixels image , Figure (4) explain the process of normalization for Arabic character(ي) [8].

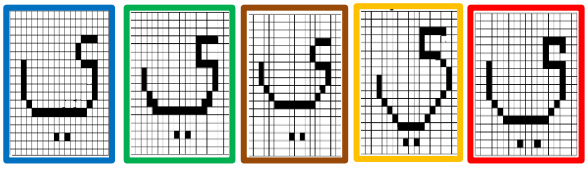
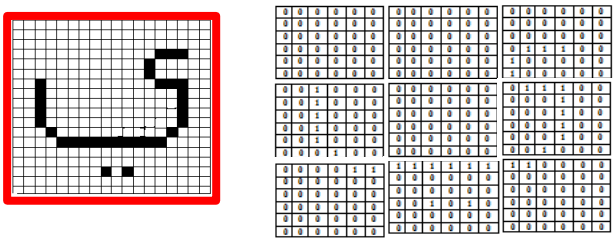


Fig. 4. Input image character after normalization

3) Feature Extraction

Feature extraction is the first step in classification and image recognition; it is the process of generating features to be used in the classification task. A method is used in our proposed neural network structure to extract features input Arabic characters based on the zoning technique [9]. The binary image is fixed and stored in a matrix form of size 18 x 18 array of binary pixels image to every Arabic character of twelve characters (ي،و،م،ك،ف،غ،ط،ش، ذ،ج,ت,ا). Figure (5a) shows the size normalized binary image of Arabic character (ي). The binary image fitted in a matrix is divided into nine square sub- matrices of size 6x6 as shown in Figure (5b) some of these sub- matrices have the value "1" when the cell, have one feature values is black and other have the value "0" when the cell is white, have zero feature values.



(A) (B)

Fig. 5. Normalized image of character "ي", (b) Partitioned into nine square regions of size 6x6

For made the feature extraction simple method for implementation, we can converted the one feature value of each sub-matrix in matrix form into a real number by the equation (5), such that the Feature value of v^{th} sub-matrix, f_v is computed by this equation. The total number of 1s cells in a sub-matrix is divided by the total number of cells in that sub-matrix [10].

$$f_v = (1/pv) \sum C_{ij} \quad (5)$$

Where

f_v= feature value of v^{th} sub-matrix

pv = total number of cells in the sub-matrix v,

C_{ij} = value of $(i, j)^{th}$ cell in the sub-matrix

The normalized feature value of v^{th} sub-matrix is computed by dividing the sum distances of 1s cells from the sum distances of all cells in that sub-matrix. The feature value of sub-matrix v is given by the equation (6).

$$f_v = \frac{\sum_{C_{ij}}^{pv} d_{ij}^v}{\sum pv\, d_{ij}^v}, \qquad d_{ij}^v = \sqrt{i^2 + j^2} \quad (6)$$

Which represent distance of $(i, j)^{th}$ cell. The set of features extracted from each sub-matrix is called feature vectors of samples which are used to train the Error Back Propagation Neural Network (EBPNN) [11].

4) Classification

Classification is performed after pre-processing step to the values of the resulting features with Error Back Propagation Artificial Neural Network (EBPANN), from this process will be obtained calculation of recognition, all the data which will then determine the percentage of success of this method [12]. The tests in our neural network consists image processing phase, feature vector extraction phase, and three network phases that is Neural network structure phase, Error Back Propagation algorithm phase, Running neural network structure

phase, and final the results phase presented in Figure (6). Our neural network structure optimized and tested 12 isolated Arabic handwritten characters (ا،ت،ج،ذ ،ش،ط،غ،ف،ك،م،و،ي) because the similarity of some Arabic characters and the location of the points in the character. four characters without pointes (ا،ط،م،و), three characters with one point above (ذ،غ،ف),one character with one point below (ج) , one character with three points above(ش), one character with hamaza (ك) , one character with two points below(ي).Each character is represented as a 18x18 array of binary pixels image, with 0 representing "off" and 1 representing "on". 252 example images were used, and these were divided into two groups: 100 images were used as training data, 240 images were used as validation data after each epoch of training data, and finally12 images were used to test the resulting network architecture. The images represented 12 isolated Arabic handwritten characters (ا،ت،ج،ذ ،ش،ط،غ،ف،ك،م،و،ي).

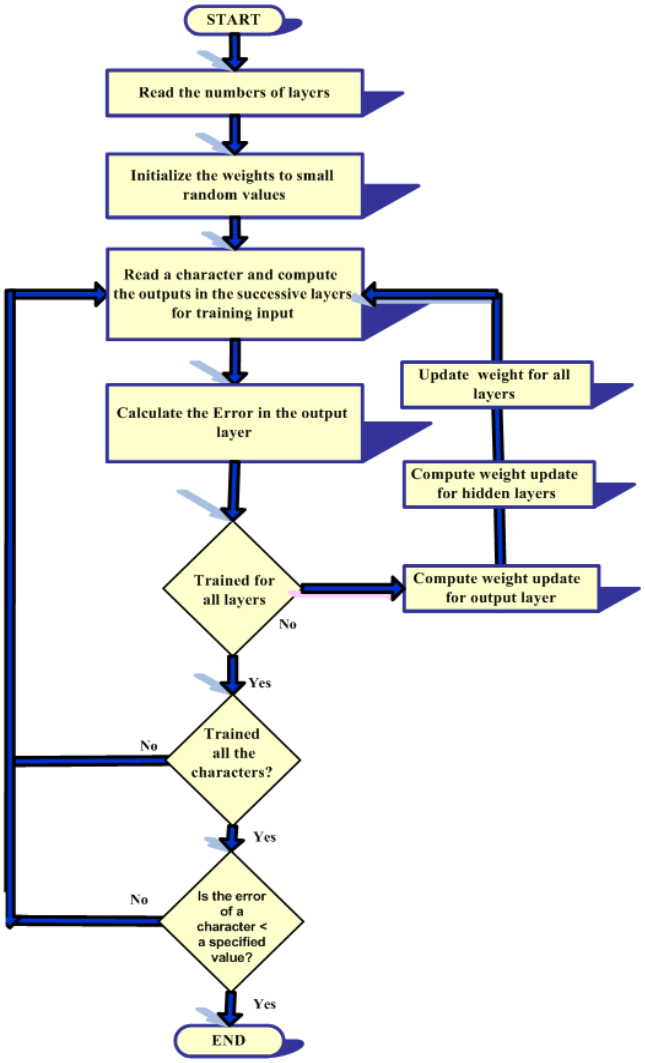


Fig. 6. The flowchart of Error Back Propagation Artificial Neural Network training architecture

IV. Knowledge Base and Experimental Analysis

A. Knowledge Base

The knowledge base contains the Arabic characters in our neural network structure. It is basically a knowledge base of isolated Arabic handwritten characters, Individual character images in the knowledge base are used to generate the values for the input character Size of each character 30x30 pixels. Knowledge base consist 12 Characters only characters (ا،ت،ج،ذ ،ش،ط،غ،ف،ك،م،و،ي) because the similarity of some Arabic characters and the location of Points in the character. four characters without pointes (ا،ط،م،و), three characters with one point above (ذ،غ،ف) ,one character with one point below (ج) , one character with three points above(ش), one character with hamaza (ك) , one character with two points below(ي),in our neural network each character has different shapes written by hand arrange in separate file for each character. The training knowledge base consists of 240 samples for 12 Arabic characters, 252characters with different shapes input samples .We use Matlab to extract the handwritten character for each file. Some training samples are shown in Figure (7).

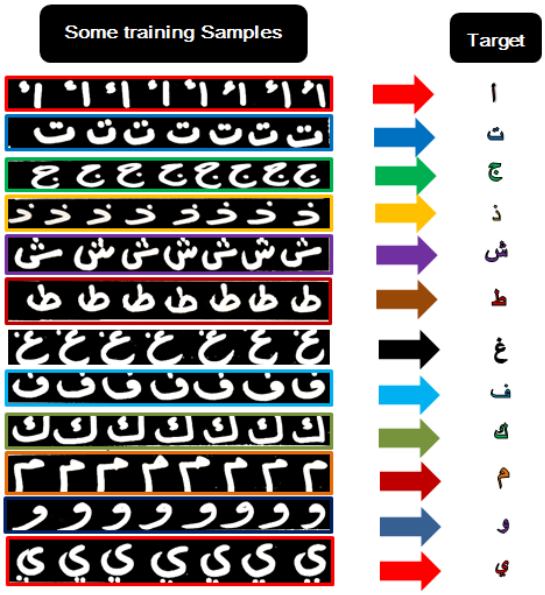


Fig. 7. Some samples of scanned offline handwritten Arabic character knowledge base

Each character is normalized into 18x18 array of binary pixels image, during preprocessing, along with noise/spurs removal. We have been extracted feature values by zoning technique for these normalized images. Thus for18x18 array of binary pixels image, number of input nodes to the neural network are 252 and learning has been implemented using Error Back Propagation Artificial Neural Network. For training purpose we have used 20 samples of each class from 12 classes and testing applied to all images. The character أ is assigned as class 1, character ت as class 2 and so on up to character ي as class 12.

B. Experimental Analysis

Experimental Analysis in our proposed neural network structure consists from three parts,

- Neural network structure.
- Error Back propagation algorithm.
- Running neural network structure.

1) Neural network structure

Our proposed neural network structure consist multi-layer, feed forward network with error back propagation algorithm is given in Figure (8). The output from one layer feeds forward into the next layer of cells, cells of each layer connect with the cells of next layer i.e. our neural network are fully connected with various weights. The obtained error is determined by the difference of outputs from output layer and target outputs. Then errors are feed backward to modify weights, which are then applied to next iteration to generate new network outputs [13].

1) Input Layers: Data sets in our this work is arranged 18x18 array of binary pixels image, the white Pixel ("off" pixel) is represented by 0, black pixel ("on" pixel) is represented by 1.

2) Hidden Layers: In our system the operation of recognition handwritten Arabic characters used three hidden layers. However, choosing the number of hidden layers accurately is deterministic in a neural network structure. If neural network contain too many hidden layers classification presentation do not agreement higher, and too little hidden layers may fail in complete high recognizing rate.

3) Output Layers: Our neural network structure applied on 12 Arabic characters for this reason it has 12 output layers sequence represent each character giving to the necessity of character recognition target.

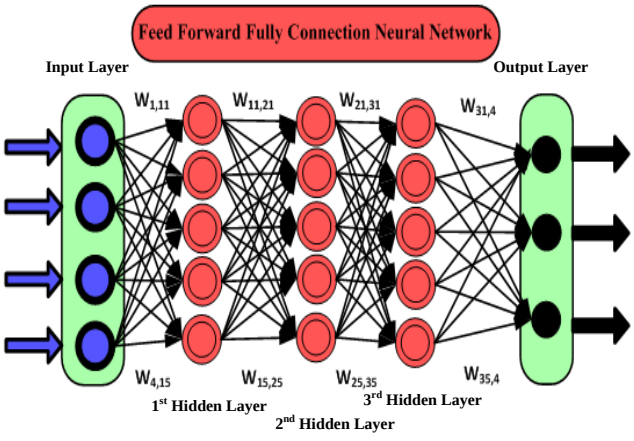


Fig. 8. Our proposed neural network structure

2) Back propagation algorithm

we give the step of Error Back Propagation Artificial Algorithm in section (4) used in this neural network structure in this section explain the equations compute the details (δ), the change of weights and updating weights is that after training stopping, In last iteration's weights should be stored as neural network weights, because it is having the lowest error which is needed to stop the training [14].

3) Running Neural Network Structure

For running the error back propagation learning process, always need set of training characters, input, and output(target), learning rate value, condition that stopping the algorithm, updating weights, nonlinearity function, in our neural network structure ,the activation sigmoid function sigmoid (bounded, monotonically increasing and differentiable) of hidden and output layers(neurons)are used. Each time all the characters in our problem have been used once in the network during training is called an epoch, small random values for Initial weight .Suppose we wanted to train a network to recognize for example the four Arabic characters (ت ،غ ،ك ,م) from 12 Arabic characters (ا،ت،ج،ذ ،ش،ط،غ،ف،ك،م،و،ي), as shown in below Figure (9).

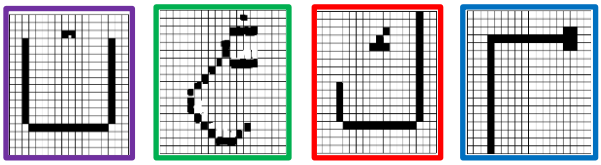


Fig. 9. the four Arabic characters (م ، ك ، غ ، ت)

To train our neural network apply the first character (ت) and change all the weights in the neural network once, which explain in Figure (5). Next apply the second character (غ) and do the same, then the third character (ك) and so on. Once we have done all four characters above, return to the first one again and repeat the process until the error becomes minim which means that it is recognizing all the characters. Figure (10) explain how our neural network should work.

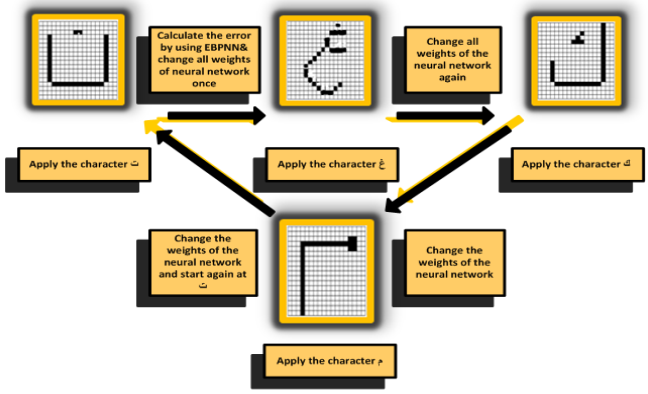


Fig. 10. Running our Neural Network Structure for four Arabic characters (ت ، غ ، ك ، م)

V. Results and Discussion

Training with validation According to the above observation, we add the restraint condition, that if error of validation is lower than 0.05, any increasing after will terminate training process. With this constrain, hopefully higher successful rate will be obtained, along with training time saving. We run the neural network for10 times, on12 Arabic characters (أ،ت،ج،ذ ،ش،ط،غ،ف،ك،م،و،ي) showing data in Figure (11) and Figure (12). The average recognition rate is 93.608%,

with average training time 36.18 seconds. Average training epoch is less than 50 iterations. Validation gives us better neural network successful rate as well as optimal training time. For each character, the neural network calculate its recognition rate, also we are showing from the Figure (13), that some characters have recognizing rate as high as 100%, but the worst case character ط can just achieve 75.520%.

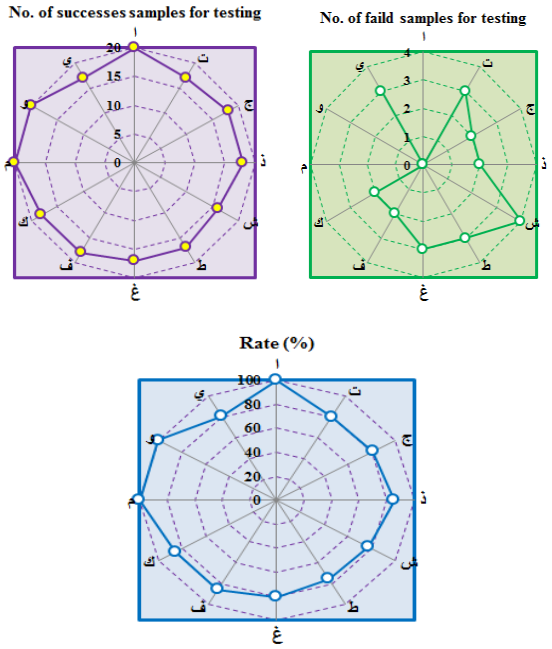


Fig. 11. Final recognition rate obtained from testing our neural network structure

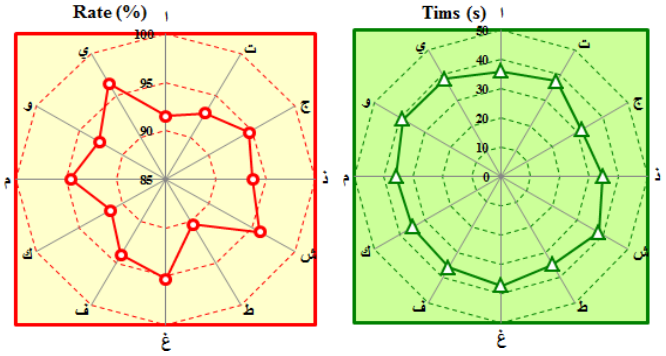


Fig. 12. Successful Classification Rate and Training Time

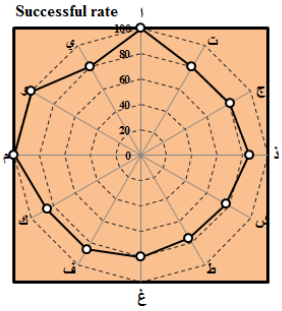


Fig. 13. Successful rate for each character using trained neural network

A. *Effect of hidden layers number*

The most commonly structure of any neural network consist of an input layer, hidden layer (layers), output layer. Number of input and output layer depend of the number of characters we want to process by our neural network and the number of characters we want to recognize and how we want to code these characters appear ,for this we must have 13 inputs characters to train our neural network because we have 12 classes for 252 Arabic characters each class contain 20 sample for every character from 12 these characters (ا،ت،ج،ذ ،ش،ط،غ،ف،ك،م،و،ي) and one class for Twelve Arabic character used for matching the recognize character when applied our neural network. for the number of hidden layers and number of their cells play important role in neural network structure because the hidden layer(s) and its (their) cells afford the network with its ability to generalize, there are no hard and fast rules for this to make the network typically works well. Theoretically neural network with one hidden layer with a sufficient number of hidden cells is capable of approximating any continuous function. Practically neural network with one and sometimes two or three hidden layers are commonly used and have well preformation. For example if we have three layer network with n input and m output cells can calculating number of hidden cells by geometric pyramid rule , described form below [15]:

$$N = \sqrt{n + m} \tag{7}$$

Where:

N: Number of cells in hidden layer.

Our neural network structure which explain in section (4.2) by Figure (8) when we want to recognize characters on 18x18 array of binary pixels image , we have 12 output cells (one for each characters from 12 Arabic characters we want to recognize). The network will train to recognize all 12 Arabic characters with anywhere between 2 and 7 hidden layer,2 layers and the network hasn't got enough weights to store all the characters, and more hidden layers need more weights in between each two layers. So updating each weight will take more time. As a result, 3-4 hidden layers may have both good performance and good converging time, above 7 layers the network becomes inefficient and doesn't perform as well. So, the number of hidden layers and its cells needs to be experimented with for the best results. We do experiments with 2, 3, 4,5,6 and 7 hidden layers, from Figure (14) , we can see that large hidden layers number leads to long training time, but did not improve identifying performance significantly. Small hidden layers number also extend the training period, and lower successful rate as well.2 hidden layers takes more than 60 iterations to converge, while the other situations having less than 40 iterations. 4, 5, 6 and 7 hidden layers have very similar rate as 3 hidden layers but takes more time to train. It is easy to understand that, more hidden layers need more weights in between each two layers. So updating each weight will take more time. As a result, 3-4 hidden layers may have both good performance and good converging time.

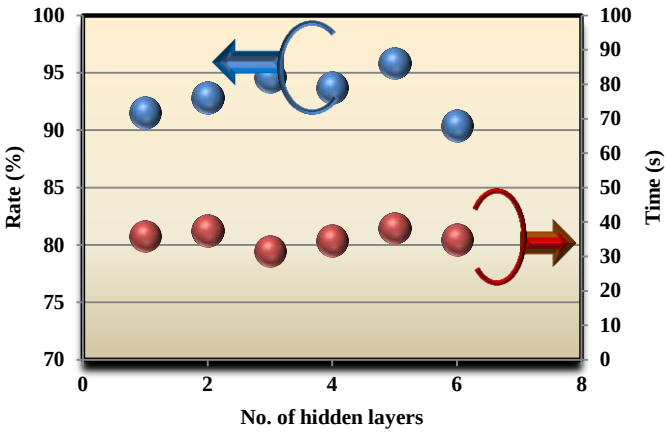


Fig. 14. Effect of hidden Layers number

B. *Effect of learning rate*

The amount of corrective modifications applied to weighs in EBPANN is called the learning rate which representation by the symbol η and is usually fixed a value between 0 and 1, so that large value of learning rate indicates large delta weights. Usually in our neural network structure choose 0.1 as the default value. And try learning value range from 0.05, 0.08, 0.1, 0.3, 0.5 and 0.7 to evaluate its effect in recognition rate and training time. After run this experiment for a number of times, we find that the minimal error of justification set is larger than 0.05, so 0.05 cannot agreement convergence. Figure (15) shows, smaller value of learning rate 0.08 has very similar act with 0.1, but large learning rate 0.3 and 1.0 will increase training time validation and decrease successful recognition rate. When learning rate equals 0.05, 0.08 and 0.1 the minimal error values are very close, also show the successful recognition rates of those three are also close. Therefore, 0.1 is appropriate rate for this knowledge base and neural network training.

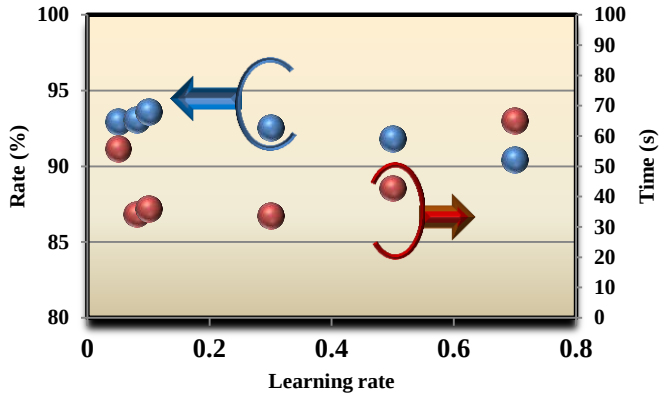


Fig. 15. Effect of Learning Rate

From testing with learning rate 0.1, hidden layers three pieces and about 252 feature vectors, this system had a level of accuracy of 100% for learning data, 80% for learning and non-learning data, and 75.520% for non-learning data

VI. Conclusion and Future Work

In this neural network structure error back propagation artificial neural network is implemented to recognize 12 offline isolated Arabic handwritten characters. As discussed above, successful recognition rate is about 93.61%, we observed in the experiments. The use of Error Back Propagation Artificial Neural Network (EBPANN) proved to be as one method to recognizer Arabic characters based on texture features with accuracy percentage at 93.61%. Determination of the learning rate value at 0.1,three hidden layers, and 132 feature vectors, all the weights are initialized to be small random number between -0.01 and 0.01, Each Arabic characters, has 20 samples for training, and testing, respectively. Can be stated that our neural network structure has been successfully calculate the learning data and less successfully for non-learning data and easy to implement and easy to compute learning rate for both hidden and output layer which modifies the values of weights and increases the convergence speed. We would like to improve the recognition accuracy percentage rate for handwritten Arabic text, which differ from the machine printed case and is more similar to cursive Latin handwritten text and cursive words or any other type of handwritten words, in our future work. By using a good feature extraction system along with other neural network classifier performs better than a single classifier.

References

[1] M. K. Mohammed Altahaf, M. B. Begum, "Handwritten Characters Pattern Recognition Using Neural Networks", International Conference on Computing and Control Engineering, 2012.

[2] Detlef Nauck. Neuro-Fuzzy Systems: Review and Prospects. In Proc. Fifth European Congress on Intelligent Techniques and Soft Computing (EUFIT'97), pp. 1044-1053, Verlag und Druck Mainz GmbH, Aachen, 1997.

[3] Nauck,D and Kruse, R., Neuro-fuzzy systems for function approximation, Proceedings of 4th International Workshop Fuzzy-Neuro Systems, 1997.

[4] Anita Pal, Dayashankar Singh "Handwritten English Character Recognition Using Neural Network" International Journal of Computer Science and Communication , Vol. 1, No. 2, pp 141-144, 2010.

[5] Sumit Goyal and Gyandera Kumar Goyal, "Cascade and Feedforward Backpropagation Artificial Neural Network Models For Prediction of Sensory Quality of Instant Coffee Flavoured Sterilized Drink". Canadian Journal on Artificial Intelligence, Machine Learning and Pattern Recognition Vol. 2, No. 6, August 2011.

[6] Zweiri, Y.H., Seneviratne, L D., Althoefer, K.: Stability Analysis of a Three-term Back-propagation Algorithm. J. Neural Networks. Vol 18, pp. 1341--1347, 2005.

[7] Efe MO , " Novel neuronal activation functions for feed forward neural networks," Neural Process Lett , Vol. 28, pp. 63–79, 2008.

[8] A. Shawkat, K.A. Smith-Miles, "Improved Support Vector Machine Generalization using Normalized Input Space," LNAI, Springer-Verlag, Heidelberg, pp.362-371.

[9] M. Kumar, R. K. Sharma, M. K. Jindal, "SVM based offline handwritten Gurmukhi Character Recognition," in Proc. of SCAKD, pp. 52-63, 2011.

[10] J. Pradeep, E. Srinivasan and S. Himavathi, " Diagonal Based Feature Extraction for Handwritten Alphabets Recognition System Using Neural Network", International Journal of Computer Science & Information Technology , Vol 3, No. 1, 2010.

[11] Ashoka H N, Manjaiah D H, Rabindranath Bera " Feature extraction Technique for neural network based pattern recognition" International journal of computer science and engineering , Vol. 4, No. 3, pp 331-339, 2012.

[12] Kundu, A., McLean MITRE Corp., T. Hines, J. Phillipsand B.D. Huyck., "Arabic Handwriting Recognition Using Variable Duration HMM." ICDAR. Ninth International Conference on Document Analysis and Recognition, 2007.

[13] Dayashankar Singh, Sanjay Kr.Singh, Dr.Maitreyee Dutta. "Handwritten character recognition using twelve directional feature input and neural network". International Journal of Computer Applications, Vol 1-No.3, pp.82-85, 2010.

[14] M. Z. Rehman , N. M. Nawi ," Improving the Accuracy of Gradient Descent Back Propagation Algorithm (GDAM) on Classification Problems," International Journal on New Computer Architectures and Their Applications, Vol. 1, No. 4, pp.838-847, 2011.

[15] S.N. Nawaz, M. Sarfraz, A. Zidouri, W.G. Al-Khatib, "An Approach to Offline Arabic Character Recognition using Neural Networks", in Proc. of 10th IEEE International Conference on Electronics, Circuits and Systems, 3, pp. 1328-1331, 14th-17th Dec. 2003.

The Parents' Perception of Nursing Support in their Neonatal Intensive Care Unit (NICU) Experience

Amani F. Magliyah
College of Public Health and Health Informatics,
King Saud bin Abdulaziz University for Health Sciences,
Saudi Arabia

Muhamamd I. Razzak
College of Public Health and Health Informatics,
King Saud bin Abdulaziz University for Health Sciences,
Saudi Arabia

***Abstract*—NICU is an environment that has many challenges in information receiving and understanding. The infants that are cared for might have serious and complex medical problems. For Parents the NICU experience is filled with stress, fear, sadness, guilt and shock of having a sick baby in NICU. The aim of this research was to explore and describe parents' experience when their infant is admitted to the NICU. And assess their perception of nursing support of information provision and according to their emotional feelings. This study was undertaken at Neonatal Intensive Care Unit in King Abdulaziz Medical City (KAMC), Jeddah, Saudi Arabia which is part of National Guard Health Affairs (NGHA) organization in the kingdom. The study utilized a self-report questionnaire with likert scale measurement and telephone interview with closed questions. One hundred and four parents agree to be the part of study and provided their consent to include their children in the study. The majority of respondents were mothers (76%), the remaining (24%) from the total sample were Fathers. All their infants have been admitted to the NICU at 2014. Many parents did not able to receive enough information easily from the unit; most of them found the information by nurses was difficult to understand. The majority of parent's perceived high stress and anxiety level according to this information. Also, Most Parents was not agreed about the nurses' support towards their emotional feeling and care. Additional finding indicate that a decrease in support level being associated with an increase in stress and anxiety level. In order to provide a high level of support and decrease the level of stress, there is a need for developing support strategies. One strategy is through a technology to develop an automatic daily summary for parent.**

Keywords—parents; stress; anxiety; NICU; nurse support; neonate; infant

I. INTRODUCTION

A neonatal intensive care unit (NICU) is specializing in the care of sick or early newborn infants [2]. Inside this unit, all of the critical life support, physiological monitoring and medical attention are provided twenty-four hours a day. The infants that are cared for might have serious and complex medical problems [6]. For parents, the NICU experience is an unanticipated journey filled with stress, emotional turmoil, strains on relationships and sometimes depression [2]. They rarely feel safe from the fear and uncertainty of the problems that can occur while the child is in care; the sequence of events in the NICU journey has many unexpected ups, downs, and turns of event [6]. Usually medical staff provides a large amount of information to the parents. Parents have to come to terms with information that they are not familiar with and have to deal with the emotional impact of the information presented to them [2]. However, the provision of clear and summarized information is important in giving parents a sense of hope and a feeling of involvement in their child's care [6]. Therefore, the research question for this study is: Are parents able to get information about their infant and received nursing support according to their feelings in their NICU experience?

Much of the literature confirms that there is a need for improving interventions, provide consistent information sharing, and high level of support for parents in the environment of NICU; to reduce parental stress, anxiety and negative feelings, improve family-centered care, and to increase parent's satisfaction and involvement in the NICU [1, 4, 5, 9]. Parents of infants admitted to the NICU will experience stress, depression, anxiety, and feelings of powerlessness, and hopelessness [9]. They used to cope with this stress via preexisting support systems of family and friends [3]. Lam et al. [4] findings indicate that parents of infants in the NICU had a lower level of stress if they perceived a high level of support, and there is a need for improving the nursing support strategies in providing parents with clear and updated information. According to Grosik et al. [2] once stressors have been identified, then interventions can be developed to improve the family-centered approach to care. It has been shown that there is a need to develop local interventions to decrease stress and enhance parents' abilities and understanding of their infant [5]. In addition to, providing holistic, family-centered, developmentally supportive care and open communication with parents in this stressful experience [9]. A 2011 study [1] found that because of an intervention the parental stress during the NICU stay was not reduced, while satisfaction with information and preference for involvement were both increased.

One means of achieving the intervention and support for parents is through developing software that generates summaries of medical data about babies in NICU using natural language processing (NLP). Research by Mahamood and Reiter [6] showed that all parents preferred texts generated with the effective strategies, and the key finding was that the use of these affective strategies might be appropriate whenever an NLP system is communicating emotional sensitive information to a non-expert recipient. Although this research provided a solution using NLP technology, it does not examine whether this software can support parents in Arab countries

and gaining their preference through generating daily summary reports. Some researchers [2, 5, 9] focused on exploring parental stressor and experience in NICU without specifying their perception of nursing support and information provision. While Franck et al. [1] measured the effect of an intervention but limited on parents stress, confidence and competence. The research by Lam et al. [4] measured the parents support provided by nurse and correlated with parents stress, it was similar to our research question but without determining the stress levels and correlated with nursing support level. Finally, these studies have not been applied in the Saudi Arabia environment. This research will present information describing the current situation in NICU parental experience, and the level of parent's perception of the information support by nurse; that will help NICU to take appropriate action and to develop an intervention to improve support strategies, promote family centered care and developmentally supportive care.

II. Objectives

NICU is an environment that has many challenges in information receiving and understanding. Parents feeling of stress, fear, sadness, guilt and shock of having a sick baby in NICU might mean that parents will not be able to process large amounts of unfamiliar information. The aim of this research was to explore and describe parents' experience when their infant is admitted to the NICU, in addition to their perception of nursing support of information provision and according to their emotional feelings. This information will help healthcare professionals to understand the parental experience in the NICU, and to develop interventions that improve supporting strategies. The following was the specific objectives; to investigate current parents experience in NICU with information provision by a medical staff; to explore the parent's ability to get needed information in NICU; and to explore parent's perception of nursing support according to their emotional feelings.

III. Methodology

The participant (parents) consent were taken along with the survey questioner in order to include in the study.

A. Study Area and Setting

This study was conducted at National Guard Health Affairs (NGHA), Jeddah. NGHA is one of the largest health organizations in the Kingdom of Saudi Arabia. The Saudi Arabian National Guard Health Affairs aim to provide the highest quality of primary, include all aspects of care and patients treatment. In addition to, the increase of awareness among members secondary and tertiary healthcare services whilst ensuring efficiency and proper utilization of available resource [7].

King Abdulaziz Medical City (KAMC) Jeddah is part of NGHA. It is offers medical care services on advance level for Saudi nationals in the Western Region ,and these services of society towards the prevention of diseases. This city makes effort to do medical initiatives and participation at all levels of local ,regional and international. Since the opening of the hospital in 1402 AH (1982 AD), strenuous efforts have been made to maintain the level of quality of care provided to patients and their advancement, until the hospital evolved and became part of the KAMC in Jeddah [3].

If a newborn baby needs intensive medical attention they are often admitted into a special area of the hospital called the Neonatal Intensive Care Unit (NICU). The NICU provides 24 hours service and only accept infants requiring level II and III specialized care. The unit is staffed with nurse to patient ratio of 1:1-2. It provides a family-centered approach to care; encompassing the parents and sick infant as a single unit. Thus including the family in all decision making and kept well informed and have complete understanding of all patient care activities/procedures. The NICU services provide the best technological evidenced – based treatment and care [8].

B. Study Subjects

The sample included both male and female parents who have an infant in NICU at KAMC hospital located in Jeddah. In order to observe and get better feeling, the sample was restricted to parents of infants who stayed one full day or more in NICU. The major key part of the sample unit for this study was on Saudi female parents between 18 and 50 years old, and Saudi male parents between 18 and 60 years old. In External validity, we can generate the sample to the target population inside the hospital, because the sample is randomly assigned. It can't be generalized to other population inside the country and to other countries due to the small sample size. Parents provided their consent in order for the children to be included in this study.

C. Study Design

The study design was a descriptive design using cross sectional study that has been carried out over a short period (4 month). It was used to estimate the participants perception of nursing support in the NICU experience and their ability to get needed information. No pre-tests and intervention were used.

D. Sample Size

The target population was all parents of infant in the NICU located in NGHA hospital at Jeddah. Estimated size of 142 parents included in the medical record at 2014 of NICU in NGHA. Normally, the both parent visit the hospital with infant thus the total population may be increased to include the other partner, but in our case we limited the population size to be as it is in the medical record and included some case where one of parents visit the hospital. The sample size was estimated and it was of 104 parents who agree to be the part of this study and provide their consent to include in the study with ±5% Precision Levels, where Confidence Level is 95% and P=.05.

E. Sampling Technique

Simple random sampling technique was used to select a representative sample of parents of infants admitted in the NICU at KAMC, Jeddah. Each participant was chosen from the medial record of NICU unit during 2014.

Since there are only 59 valid records, the rest participants were chosen when they come to see their infant or for their infant's vaccine. The process continued until sufficient participants have been identified to meet the desired sample size.

F. Data Collection methods, instruments used, measurements

The researcher aimed to identify parent's self-reported opinion about their perception of nursing support and to describe their NICU experience with current and traditional information provision. The questionnaire survey contained three parts; the first one was the personal basic information which is name, age, gender and educational level and their consent to include in the study. The second part about the parents experience with information's by nurses in NICU. Last part was about the current nursing support of their negative emotional feelings.

TABLE I. PARENT CHARACTERISTICS (N=104)

characteristic		n (%)	Mean	Std. Deviation
Gender	Male Female	25 (24%) 79 (76%)	-	-
Age	<20-30 >30-50	48 (46%) 56 (54%)	30.53	6.49
Education Level	1. ElementarySchool 2.HighSchool 3.UniversityDiploma 4.UndergraduateUniversity 5.PostgraduateUniversity	21 (20.2%) 39 (37.5%) 1 (1%) 41 (39.4%) 2 (1.9%)	2.65	1.24

The survey was tested, and all questions were revised and it was found that, the personal information part had no modification. The Questions in second part were found to be a little bit confusing, these questions were retyped and rephrased to be clearer. In addition, response categories in questions eight and nine were modified to be in a form of measuring their current feeling in NICU experience when nurses told them information about their preterm infants. Last part was clear to them and no modification was applied. The question number five, ten and twelve in the survey will be used to collect the research variables.

The researcher used a telephone interview that requires the respondent to answer the closed question in the questionnaire. In addition to a self-administered survey technique that requires the respondent to complete the questionnaire by him/herself. These surveys were distributed in-person and responses were collected directly into the questionnaire, which needed to establish a good relationship with participants (parents) and a comfortable flow of questions, in order to ensure that the appropriate data are collected. The researcher wanted to reach a sample size of 104 parents, and assumed that the response rate might be 95% of the needed sample size. Initial sample size was parents of infants at NICU with lower levels of specialized care at KAMC hospital in Jeddah.

G. Data Management and Analysis

Data were tabulated on IBM SPSS sheet and have been analyzed using the statistics in the SPSS software for Windows. Discrete data were analyzed using use CI 95% for the descriptive data to measure the mean and standard deviation of the scores.

A convenience sample of 104 parents agreed to be part of the study and provided their consent in order for the children to be included in the study. Sixteen parents answered to the questionnaire's questions by a phone interviews. And forty four parents answered the questionnaires via self-administered method. The majority of respondents were mothers (76%), the remaining (24%) from the total sample were Fathers. All their infants have been admitted to the NICU at 2014.

The characteristic description of the parents is shown in Table 1; the mean age of the participant was about 31 years old; the majority of them were between 30 to 50 years; in the education level most of them had an undergraduate university degree (39%) from both genders, and a high school level (38%) where most of them were mothers; 20%had an elementary school level and all of them were mothers; 2% of female persons have postgraduate university degree, and only 1% of males had a university diploma.

The parent's perceived levels of agreement associated with information's provided by Nurse in NICU in addition to the levels of stress are shown in Table 2. The mean score for each perceived rates is shown. The range of scores was from 1-5 (from positive to negative) that indicate Likert scale level of measurement and agreement. The findings present that a total of 51% of parents disagreed that they can easily get information and help from nurse when they visited or telephoned the unit, while 40% parents are agreed. 5% of them rated neutral and 4% are strongly disagreed. Most of parents (64%) are disagreed that they received from nurses enough information daily about their baby progress, however 18%of parents agreed on that. 9% of them rated neutral and 9% are strongly disagree. The percentage of parents who disagreed that they understood what nurses told them about their infants is 59% giving the most rates, just 23% are agreed and 1% was strongly agreed. 10% of them rated neutral and 7% are strongly disagreed.

TABLE II. MEAN AND STD. SCORE RELATED TO INFORMATION'S PROVIDED BY NURSE IN NICU

	Mean	Std. Deviation
Easy to get information or help from nurses	3.18	1.02
Received enough information daily about my baby progress	3.63	.88
Understood what nurses told me about my baby	3.47	.95
Stress after nurse's informations about my baby's condition	1.88	.37
Anxiety after nurse's informations about my baby's condition	1.79	.49

In the NICU environment the infant's status can be changed frequently, many tests can be performed and may include many treatments. The majority of parents (85%) had a high stress level after nurses told them about their infant's condition, while 13% of them had a very high stress level and only 2% had a moderate level. Most of parents (71%) had a high anxiety level after nurses told them about their infant's condition. While 25% of them had a very high anxiety level and only 4% had a moderate level.

The parent's perceived levels of agreement associated with nursing support of parents feelings during the NICU

experience are shown in Table 3. As previous part the mean score for each perceived rates is shown. Most of parents (68%) are disagreed that nurses responded to their worries or concerns clearly and 1% is strongly disagreed. 20% of them rated neutral and 11% are agreed. A total of 73% of parents disagreed that nurses allowed them to be involved in their infant's care, while 21% are agreed. 4% of them rated neutral and 2% are strongly disagreed. The last one was a bout nurse' support in general towards parents their infant's hospitalization. 60% parents are disagreed, that means the support level was low and 1% is strongly disagreed that indicate very low support level. 35% of them rated neutral and only 4% are agreed and it indicates that they had a high support level.

TABLE III. MEAN AND STD. SCORE RELATED TO EMOTIONAL SUPPORT PROVIDED BY NURSE IN NICU

	Mean	Std. Deviation
Nurses responded to my worries or concerns clearly	3.60	.69
Nurses allowed me to be involved in my baby's care	3.56	.84
Nurses' support towards me during my baby's hospitalisation.	3.58	.58

From the results we found that most parents (n=53, 51%) who disagreed about the given nursing support, received a high stress and it is most percentage in the stress level as shown in Table This means a decrease in support level being associated with an increase in stress level. However, the intersection of the parents who received a high anxiety level and disagreed about the given support was 44% of them (n=46), and it is lower than the stress. In addition there were some parent's perception of stress and anxiety did not match with their perception of support.

IV. DISCUSSION

The results of this study have shown that parents of infants admitted to the NICU had a high level of stress, high level of anxiety and a low level of support from nurses according to infant's information. This is an interesting finding that most parents find the experience in the NICU stressful. This study is consistent with earlier research [4]; they found a moderate correlation between stress and support variables, which indicates that "with a high level of nurse support, the stress perceived by the families would be less". Another previous study [2] found that the highest stress scores were in parents observations for their infants in distress or when appearing very ill. Also parents' were most stressful by the noise of alarms and machines located in the unit. Furthermore, they proved that parents were struggle of separation and helplessness from their babies in the NICU [2]. However, Linda & Trudi [5] have shown that parents were experienced a moderate to high stress levels regarding their infant's stay in NICU. It has been clear that "having an infant in the NICU was an overwhelming experience associated with negative feelings. These included role strain, distress, and emotional pain" [9].

TABLE IV. P.STRESS * NURSESUPPORT CROSSTABULATION, SPSS

Count

		NurseSupport				Total
		Agree	Neutral	Disagree	StronglyDisagree	
P.Stress	VeryHigh	1	3	9	1	14
	High	2	33	53	0	88
	Moderate	1	1	0	0	2
Total		4	37	62	1	104

Many changes could occur frequently in the infant's condition during hospitalization such as surgery and post-operative processes [5]. This study showed that the majority of parents didn't understand the information provided by nurse. In the study done by Linda & Trudi [5] shown that parent want to be understood by medical staff in the unit, not just by their friends and family. And in their situation they need for empathy to them not a pity from people. In addition, "They wanted clear and accurate information, guidance, understanding and empathy from people around them" [5]. Parents were considered the infant pain as very essential value and they wanted more of information to be involved with their infants, especially during these painful procedures [1].

Lam et al. [4] found that parents who didn't understand their baby's condition were perceived as highly stressful experiences. Also, they found that "some parents did perceive that nursing communication and behavior contributed to their stress, in particular when they felt the nurses did not give them enough information about tests or treatments being undertaken" [4]. This correlate with our result which found some of parents didn't able to get information about their baby's condition and progress easily from NICU. In addition most of them didn't get enough information daily from the nurses. Regarding to these findings lam et al. [4] had mention that in order to help the families cope, the staff in a particular NICU have to develop several strategies to use it in their practice. Linda & Trudi [5] found that the communication was an essential aspect that parents care about in their experience, in which that the ways of perceiving the communication wither in a good or bad way had impact their stress experience directly. While the communication was recognized as important, parents were needed to understand their infant's status and progress, and the care delivered to them. In addition, "The manner in which staff went about their business was recognized as a potential stressor" [5]. That means when a good communication and education were provided by staff, and parents were involved in the care, this had impact positively on the parents' experience. But when the staff practices or behaviors are inconsistent this was considered as unprofessional and the parental stress level was increased [5].

Our result shown that parents' perception of receiving support from nurses was low, nurses didn't respond to the worries and concerns of most of the parents. Lam et al. [4] said that there is a need for an emotional and psychological support to take in the consideration, and nurses have to provide a level of support to meet with parents need.

In addition, our study finds that nurses didn't allow parents to be involved in their infant's care." Both mothers and fathers spoke of stress from the separation from their baby. This experience was intensified when the baby was unstable, whilst the mother was in the postnatal ward and for mothers when breastfeeding" [5]. Obeidat et al. [9] found that the mothers was needed to be closed ,near and belong to their babies, and if these needs were achieved, the mothers will have a feelings of responsibility, confidence , and familiarity with their infant . Furthermore, "when parents were involved in infant care, were allowed proximity, communicated clearly and openly, and formed rapport with the nurses, they became more satisfied and confident in their parenting roles"[9]. A previous study by Franck et al. [1] have demonstrated that most of parents were realized that the intensity of their infant's pain was a little high, thus their preference to be involved actively during these procedures was very strong. Also they found that "mothers who participated in a Newborn Individualized Developmental Care and Assessment Program reported feeling closer to their infants but also experienced a higher level of anxiety than mothers who did not participate in the program" [1].

In this study, these findings has leads the researchers to consider a technical way of to develop and provide support for the parents. The reason for thinking about technology rather than other strategies is the natural of NICU's environment." Sometimes the busy workload may contribute to a perceived lack of time available for supportive care, or nurses may even forget that parents are heavily dependent on them for emotional support and parental guidance" [4]. Mahamood and Reiter [6] had presented a system with affective NLG strategies for medical texts generation that developed for parents of pre-term infants. These generated texts are an English summary contains data and information about infants in a NICU, and it serves as stress reduction tool. They found that most of parent preferred using the system, they found it emotionally appropriate, helpful, and the information was understandable [6]. Thus, the generation of such reports for parents was useful and successful. So, further research on technical ways is required in order to increase parental guidance, support and involvement, to achieve parents' satisfaction during the neonatal stay, especially in Arabic countries.

V. LIMITATION

In the study where the researcher analyzed and explored the perception of nursing support to parents in NICU experience, the main limitations were the small sample size and the cross-sectional study design. Also some of data collection process conducted through surveying the participants in one and busy NICU unit during a short period. In addition, there were a number of threats to internal validity. First, selection bias was likely to be a more significant threat, participant may have been chosen intentionally, or it may not have been possible to randomly assign participants. To avoid this type of a threat, authors used random assignment for Parents' selection so that they were being measured across a range of general and specific criteria such as age, behavior, gender, educational level and morality. Second, the participant reactivity and behavior; most of the parents were tired, depressed, bored, not cooperative and inattentive. Such factors were difficult to control and it could reduce the internal validity. Thus, the authors used mixed method for collecting data and add phone interview method; because the parents feel more comfortable than when they are in the hospital. Also mixed methods used to reach the desired sample size in the small period.

Third, statistical regression was a threat; the scores of individuals on the dependent variable were not only being due to the natural performance of those individuals, but also measurement errors or chance. In addition, the researcher effect that is typically unintentional, but arises because of the personal characteristics of the researcher that influences the choices made during a study, and non-verbal cues that the researcher gives out that may influence the behaviour and responses of participants. Finally, the testing threat was controlled because this was not an experimental study and there was no pretest. Also changes related to time (maturation, history etc.) were not measured because there was no pretest.

VI. RECOMMENDATIONS

The finding from this research is evidence that we need a solution to improve parents' perceptions about neonatal experience. The authors recommend to develop Automatic Summary Generator System in order to Supporting Parent's at NICU , reduce the stress and anxiety level to an acceptable levels, and provide a support to parent to increase their satisfaction about the support delivered in this environment. The system used natural language processing technique (NLP) and it has used a lot in the creation of e-Health systems especially in generating information [6]. It assists healthcare practitioners in providing information support to doctors and nurses [6]. Beside this, NLP technique is "playing a greater role in providing patients with access to information in a personal form"[6]. In order to implement this system, the authors suggest following the study conducted by Mahamood and Reiter [6], begin with the management plan, then the design, implementation, and evaluation of this system. It have to be adapted it to be suitable to Arabic parents in the Kingdom.

VII. CONCLUSIONS

One hundred and four parents agree to be the part of study and provided their consent to include their children in the study. The majority of respondents were mothers (76%), the remaining (24%) from the total sample were fathers. This study showed a high level of stress and anxiety and a low level of support have been perceived by parents in NICU experience. It helps to recognize the importance of establishing support with parents during their infant's hospitalization in the NICU where most of negative feelings occur. The study has also shown that nurses did not provide emotional support to parents. A recommendation from the findings of this study presents that there is a need for an Automatic summary generator for parents to alleviate their stress, anxiety, increase support, and care involvement, as a support strategy that can improve parent's emotional support with care and information.

ACKNOWLEDGEMENT

The author thanks Ms. Intisar Abdullah, MS.RHIA, and Director Health Information Management Department at NGHA Hospital for her support in data collection through

providing parent's phone numbers in medical records. This work has been financially supported by King Abdullah International Medical Research Center (KAIMRC) National Guard Health Affair, Saudi Arabia through grant SP14/059.

REFERENCES

[1] Franck, L. S., Oulton, K., Nderitu, S., Lim, M., Fang, S., & Kaiser, A. (2011). Parent Involvement in Pain Management for NICU Infants: A Randomized Controlled Trial. Pediatrics , 128 (3).

[2] Grosik, C., Snyder, D., Cleary, G. M., & Breckenridge, D. M. (2013). Identification of Internal and External Stressors in Parents of Newborns in Intensive Care. The Permanente Journal. , 17 (3), 36-41.

[3] King Abdul Aziz Medical City - Jeddah. (n.d.). Retrieved from NGHA: http://www.ngha.med.sa/Arabic/MEDICALCITIES/Jeddah/Pages/default.aspx

[4] Lam, J., Spence, K., & Halliday, R. (2007). Parents' perception of nursing support in the neonatal intensive care unit (NICU). Neonatal, Paediatric and Child Health Nursing , 10 (3).

[5] Linda, S., & Trudi, M. (2012). Identification of parental stressors in an Australian neonatal intensive care unit. Neonatal, Paediatric and Child Health Nursing , 15 (2).

[6] Mahamood, S., & Reiter, E. (2011, September). Generating Affective Natural Language for Parents of Neonatal Infants. Association for Computational Linguistics , 12–21.

[7] National Guard Health Affairs Hospitals. (n.d.). Retrieved from Geneva Health: http://www.genevahealth.co.uk/international-opportunities/middle-east/working-in-saudi-arabia/national-guard-health-affairs-hospitals.aspx

[8] Neonatal Intensive Care Unit . (n.d.). Retrieved from NGHA: http://www.ngha.med.sa/English/MedicalCities/AlMadinah/NursingSvc/UnitsServices/Pages/NeonatalIntensiveCareUnit.aspx

[9] Obeidat, H. M., Bond, E. A., & Callister, L. C. (2009). The Parental Experience of Having an Infant in the Newborn Intensive Care Unit. The Journal of Parental Education , 18 (3), 23-29.

Hybrid PSO-MOBA for Profit Maximization in Cloud Computing

Dr. Salu George
Assistant Professor
School Of Information Technology
Al Dar University College, Dubai - UAE

***Abstract*—Cloud service provider, infrastructure vendor and clients/Cloud user's are main actors in any cloud enterprise like Amazon web service's cloud or Google's cloud. Now these enterprises take care in infrastructure deployment and cloud services management (IaaS/PaaS/SaaS). Cloud user 's need to provide correct amount of services needed and characteristic of workload in order to avoid over – provisioning of resources and it's the important pricing factor. Cloud service provider need to manage the resources and as well as optimize the resources to maximize the profit. To manage the profit we consider the M/M/m queuing model which manages the queue of job and provide average execution time. Resource Scheduling is one of the main concerns in profit maximization for which we take HYBRID PSO-MOBA as it resolves the global convergence problem, faster convergence, less parameter to tune, easier searching in very large problem spaces and locating the right resource. In HYBRID PSO-MOBA we are combining the features of PSO and MOBA to achieve the benefits of both PSO and MOBA and have greater compatibility.**

Keywords—Cloud Computing; Profit Maximization; Admission Control; SLA; Optimization; Hybrid Particle Swam Optimization – Multi Objective Bat Algorithm

I. INTRODUCTION

Cloud Computing is business enterprise which invests its capital in deployment of infrastructure for developing data centers. Therefore, profit is the main aim of any enterprise and maximizing its profit on the large investment is of the major research contribution in any business. Since cloud computing is going to bring a greater revolution in the real life it needs to concentrate in its gain and sustain in the competitive market by providing computing at low cost without scarifying the quality of service and as well as maximize the cloud service provider 's profit.

Cloud Computing made the dream of computing become true. Cloud computing represents real paradigm shifts in the way in which system are deployed. The massive scale of cloud computing systems was enabled by the popularization of the internet and the growth of some large service companies. Cloud computing makes long-held dream of utility computing possible with a pay-as-you-go, infinitely scalable, universally available system. With cloud computing you can start very small and become big very fast. That's why cloud computing is revolutionary, even if the technology it is built on its evolutionary.

Cloud Computing is business enterprise which invests its capital in deployment of infrastructure for developing data centers. Therefore, profit is the main aim of any enterprise and maximizing its profit on the large investment is of the major research contribution in any business. Since cloud computing is going to bring a greater revolution in the real life it needs to concentrate in its gain and sustain in the competitive market by providing computing at low cost without scarifying the quality of service and as well as maximize the cloud service provider 's profit.

II. PROFIT MAXIMIZATION IN CLOUD

Profit maximization is a process by which a firm determines the price and output level that returns the maximum profit [3]. Any firm must ensure that it provides the products at cheaper rate along with quality. They should not provide low quality and user satisfaction is important. Cloud computing takes technology, services and application on remote system through Internet, into self service utility. Computing as utility is one of the biggest business today IT is revolving to and apart from providing services at low these firms are interested in gaining the maximum profit on large investments made on datacenter of cloud [6].

We consider Firm's behavior on profit maximization-Firm's scale decision which implies Firm maximizes the profit by considering cost function which minimizes the cost and the optimal quantity of given market prices. Cloud computing as discussed earlier convert's capital expenditure to operational expenditure, and provides on demand access to pool of resources. So, Cloud service providers face many challenges to create successful business apart from which they need to retain profit. Many cloud providers are vague and need to sustain in perfectly competitive cloud computing market and maximize their profit. Cloud computing economics usually considers either SLA agreement or Resource Scheduling policies. In this discussion we consider Infrastructure Vendor, Cloud Service Provider and Customer economic Strategies.

A. Profit Maximization Factors

1) Infrastructure Vendor: to retain profit an Infrastructure Vendor's economy of scale should consider:

- Low cost electricity
- High network bandwidth pipes
- Low cost commodity hardware and software.

Apart from the above mentioned it also need to consider the usage pattern, types of request and infrastructure costs [7][20][21][22].

2) Cloud Service Provider: A cloud service provider is one who builds the communication model of cloud where infrastructure vendor is cloud storage model. It is responsible for user satisfaction –important feature of any business where cloud provides business agility(maximizes returns by distributed parallel programming), and IT Efficiency(minimize cost through virtualization).To maximize the profit cloud economies should consider :

- SLA agreement satisfying users and maximizing profit.
- Cost of renting (type of service).
- Configuration of multiserver system.
- Resource scheduling.(profit aware and suitable for distributed parallel computing).[8][9][10][11]

3) Customer/Cloud User: Customer should take few responsibilities of

a) Job should be completed on time with minimum cost.

b) Characteristics of application.

c) Amount of a service.

Cloud profit maximization has considered study only on either infrastructure vendor or service provider with any one issue (SLA, Resource Scheduling etc or few of them). But we here put forth all the necessary factors to be considered while maximizing the profit in a cloud[16][17][18][19].

B. Measuring Cloud Computing Cost

An application in cloud computing provides economics of scale commodization of assets, and conformance to programming standards. The cost of a cloud computing deployment is estimated as

$$\mathbf{Cost_{CLOUD}= \sum(UnitCost_{CLOUD}*(Revenue-Cost_{CLOUD}))}$$

where unit cost=cost of a machine instance per hour or another resource[2][3].

Cloud enterprise makes computing not only cheaper but also faster and efficient. Thus, cloud computing takes all the hidden barrier's upon and comforts the end user. The demand thus, is more and it is measured in terms of Compute Unit or CU. It always has expands and shrinks upon the demand , hence Right-sizing is value of proposition in cloud computing cost. The data center needs to be fully utilized and systems being ideal should be utilized rightly to charge effectively and maximum not only returns but maximum returns. Optimization has always been a major concern in any resource utilization and thus our ideas follow them effectively.

Operational cost considers ROI (Returns on investment) as its metrics to calculate Total Cost of Ownership (TCO). Due to cloud flexibility and agility today capital expenditure is converted to operating expenditure [2][3].

Hence, to measure a cloud computing cost:

1) Customer should provide:

- Type of service need and levels of services (if required).
- Amount of storage required.

2) Service Provider:

- SLA which provides quality of service, satisfaction of customer and penalty charges.
- Risk and Uncertainty.
- Optimized Resource Scheduling.[1]

3) Infrastructure Vendor:

- Cost of renting (low cost commodity hardware or software).
- Cost of energy consumption (low cost electricity).
- Cost of network.

C. Resource Allocation and Optimization

Various resource scheduling strategies have been implemented in cloud environment. Proper resource Scheduling is to utilize the resources efficiently. In High Performance Computing where multitenant system is needed optimization is best technique to increases the profit. In cloud demand spikes should be handled in order to sustain inn competitive market among different cloud providers. Cloud computing should therefore optimize resource usage.

Cloud environment uses probabilistic algorithm, Monte-Carlo algorithms of which prominent are 'Ant Colony Optimization', 'Particle Swarm Optimization' and 'Genetic Algorithm.

Cloud for its maximum returns considers many objective such as minimizing cost and optimizing pooled resources in VMs. To rent resources at low cost it has to utilize the resources efficiently. So, we propose Bat intelligence which uses multi objective optimization.

High performance Computing needs parallel programming in distributed environment. Bat Algorithm is proved its global Convergence which is not achieved in any optimization. Thus, this type of resource allocation ensures maximum profit. Our goal is to achieve maximum profit without any sacrificing of services provided by cloud efficiently and effectively [1] [16] [21] [23].

III. PROPOSED PROFIT MAXIMIZATION FRAMEWORK IN CLOUD COMPUTING

The objective of the project is to maximize the profit in cloud which needs to consider, Resource scheduling, Power consumption model, QOS (Risk and Uncertainty), Optimal speed and Size of servers. In order to overcome these problems proper admission control, resource scheduling and optimal multiserver configuration is to be performed. To maximizing the profit the cloud has to take profit maximization technique which minimizes the cost and optimizes the profit. Therefore Cloud service provider has to manage and optimize the pool of resources along with the cloud user satisfaction and within the given infrastructure to its deployment. Cloud Computing provides computing as a

Utility. Like measuring any other utility (such as electricity, water, etc) computing should only charged as per usage i.e. Pay-as-you-go on basis of cloud.

A job is submitted to cloud by a client. The job or request first enters the queue. To avoid unexpected loss we need to concentrate on risk and uncertainty by controlling the admission of job and finally once the SLA is signed the optimized resource allocation is done which will compute fast with efficient processing. Once the job is assigned to VMs it should start the process and complete as per SLA.

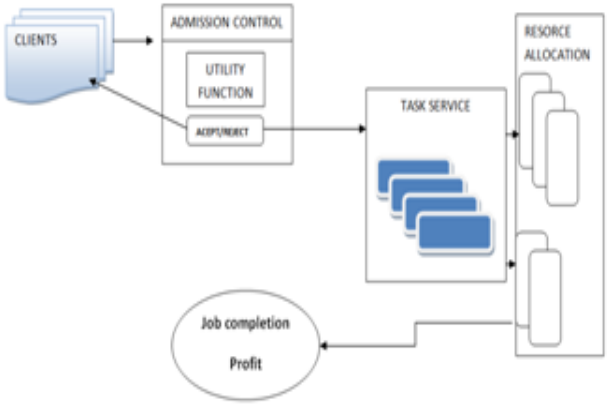


Fig. 1. Proposed Framework for Profit Maximization in Cloud Computing

A. Profit Aware SLA Specification

In cloud computing SLA is service based as it focuses on characteristics of datacenter and network to support end-to-end communication. If the response time to complete a task is less than specified time in SLA then the service is best and gets its credit. The credit again can be more profit aware if it considers uptime and double the uptime (which implies double the payment i.e. best QOS). But if the response time is longer than the specified time in SLA then penalty is induced (reduce the cost), for low quality of service. And if still it prolongs the response time beyond the limit (waiting to longer) than the cost incurred is zero as per SLA specified. Therefore we usually go with Hard SLA where violation will not reflect profit.

B. Admission Control

Profit aware cloud considers utility function which is price that the customer is willing to pay. If the job can completed within specified response time than the job is accepted but a service provider is also doing business where gain cannot only relay on admitting possible things but retaining its customers in a competitive cloud market is important as rejection will lead to loss of business. Therefore, a service provider should consider yield function with weight functions of cost minimizing(CPU and time) and trustworthy customer(Cloud Client Register with service provider) to sustain in the market for long (i.e. regular at payments).

C. HYBRID PSO-MOBA

We propose a HYBRID PSO-MOBA, (Hybrid Particle Swarm Optimization and MultiObjective Bat Algorithm) which combines PSO(Particle Swarm Optimization) and MOBA(MultiObjective Bat Algorithm). PSO search takes place in local space and global updation is done by MOBA. Cloud user submits the job to service provider where we take the probabilistic optimization of all jobs in queue through M/M/m queuing model. Then we take multiobjective resource allocation to services in by MOBA, where the first objective is to consider cloud user's and then client willing to pay double and finally we need to consider the expected service time. The process updates in global search space through MOBA.

Hybrid PSO-MOBA.

Initialization:

Initialize a population array of particles with random positions and velocities on *D* dimensions

in the search space.

Iterative loop:

For each particle, evaluate the desired optimization fitness function in *D* variables.

Compare particle's fitness evaluation with its *pbesti* . If current value is better than

pbesti, then set *pbesti* equal to the current value, and _ *pi* equal to the current location

_xi in *D*-dimensional space.

Identify the particle in the neighborhood with the best success so far, and assign its

index to the variable *g*.

Repeat:

Change the velocity and position of the particle according to the following equation

_

_vi ← *_vi* + _ *U*(0,*φ*1)⊗(_ *pi* − *_xi*)+ _ *U*(0,*φ*2)⊗(_ *pg* − *_xi*),

_xi ← *_xi* + *_vi* .

Until a complete schedule is constructed

Apply MOBA search process

Apply the global updating rule

If a criterion is met (usually a sufficiently good fitness or a maximum number of

iterations), exit loop.

end loop

MOBA search process

Objective function $f_1(x)$, …….$f_k(x)$,$x=(x_1,….x_d)^r$

Initialize the bat population $x_i(i=1,2,……,n)$ and v_i
For j =1 to N
Generate K weights $w_k>=0$ so that $\sum^K_{k=1} w_{k=1}$
From a single objective $f=\sum^K_{K=1} w_k f_k$
while(t<Max number of iterations)
Generate new solutions and update by (1) to (3)
If(rand>r_i)

```
  Random walk around a selected best solution
      End if
      Generate a new solution by flying randomly
     If (rand<A_i&f(x_i) < f(x*))
  Accept the new solutions,
  and increase ri& reduce Ai
end if
     Rank the bats and find the current best x∗
end while
 Record x∗ as a non-dominated solution
end
Postprocess results and visualization
```

Algorithm: Hybrid PSO-MOBA

1) Cloud user

Cloud user needs to provide the amount of CPU, memory and time and factors to maximize the profit considered are a) amount of a service (requirement of a service r) and the workload of an application (λ). Cloud user shouldn't request for more resources than needed. Our approach checks for regular cloud customer, service charge (the client is willing to pay) and execution time (within average execution time implies no penalty).

2) Admission Control

Admission control is responsible for determining whether it will be profitable for the service provider to accept a job.

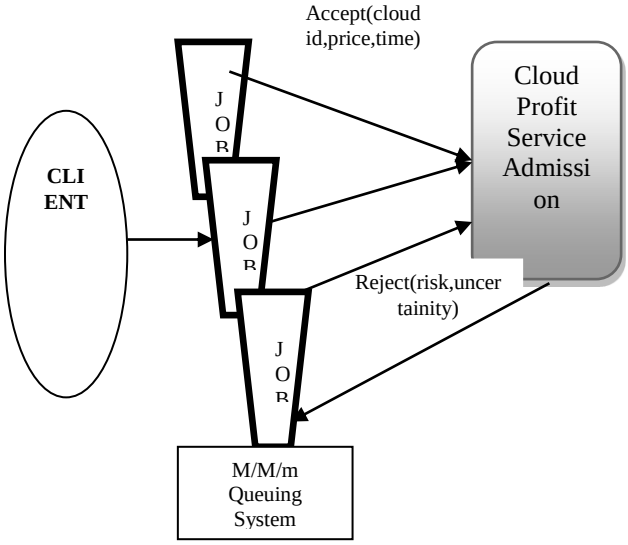


Fig. 2. Admission control flow

Profit-aware control with the time, price, client constraints to utility function. The tupple in utility function is represented by, Utility (risk factor, time, cloud user).Input for the admission control are a) cloud id (regular cloud customer), b) Client willing to pay and c) Expected service time. Output from admission control i.e. the accepted is sent to scheduling and optimization pool and rejected is sent to the user request to be accepted later or released. Advantage of Admiting process is it makes scheduling easier and profitable (economical oriented).

3) Resource scheduling

Appropriate resource Scheduling is necessary to configure a multiserver system. We consider scheduling as NP Hard problem and select the best scheduler algorithm micro bat algorithm. We propose a profit aware bat algorithm which optimizes the resource allocation depending upon the service charge and business cost .Since bat algorithm is the only algorithm which has good performance of global convergence property. Input given is the independent jobs with priority which to complete first depending upon admission. To obtain the best optimized resource allocation we have HYBRID PSO-MOBA.

4) To maximize the profit

When the server size is small the waiting time of the request is long also gain and the service charge is low. If the server size increases the waiting time will become decreased also the service charge and the gain are increased. It is also applicable to the speed of the server. Waiting time of the request is increased, gain and the service charge is decreased when the server speed is low. Waiting time of the request is decreased, gain and the service charge is increased when the speed of the server is high. Cloud service provider manages the multiserver System through M/M/m Queuing System and optimizes the resource using Hybrid PSO-MOBA. Resource allocation according to admission control and profit aware SLA. To Provide efficient multiserver (powerful) than more servers as it will increase cost of renting and power consumption. General purpose optimization algorithms are simply not suitable for solving this kind of puzzle. EA Scheduling proprietary algorithms, coupled with advanced heuristics, deliver highly optimized schedules blisteringly fast—even for the largest and most complex scheduling problems.

D. ADVANTAGES

Service provider allocates resources and schedules tasks in such a way that the total profit earned is maximized. Our methodology can be applied to other pricing models. The cost of the service is also reduced. Managing Multiserver through M/M/m Queuing model will provided efficient resource management. Optimizing Resources through MOBA will achieve Faster Global convergence. This effective resource optimization will lead to best utilization of resources and as well as Effective and powerful server which ultimately maximizes the profit.

IV. CONCLUSION AND FUTURE ENHANCEMENT

Optimization of cost model, agility and scales are primary value proposition of adopting a cloud computing based on pay by use, scalable infrastructure and platform services. Organization need to analyze their application portfolio to profile applications which would be adaptable for cloud computing models. The Cloud infrastructure once setup is business investment which needs to return maximum profit over the time period for which we need to consider mainly low power consumption, high performance computing, optimized resource allocation with SLA policy satisfaction.

In future, we need to develop different factors of VM allocation, energy efficiency, different levels of service, and cost of network. We need to develop individual factor based

profit system, but considering all factors ensures no loss in business, which itself ultimately leads to profit. A perfect framework with all the factors with market oriented approach is to be enhanced.

REFERENCES

[1] R.RAJU ,M.Kalaiarasi, "Profit Maximization in Cloud Computing", accepted for presentation and as well as publication in IEEE digital library,at IEEE- International Conference on Information Communication and Embedded Systems "ICICES-2014".

[2] M. Armbrust et al., "Above the Clouds: A Berkeley View of Cloud Computing," Technical Report No. UCB/EECS-2009-28, Feb. 2009.

[3] Barrie Sosinky,"Cloud Computing Bible",Wiely publication,2011.

[4] http://www.infoworld.com/d/cloud-computing/when-it-makes-sense-become-cloud-provider

[5] http://en.wikipedia.org/wiki/Profit_maximization,2013.

[6] http://home.uchicago.edu/~vlima/courses/econ201/pricetext/ProfMax.pdf.

[7] Calheiros, R.N. , Thulasiram, R.K. ,Buyya et al, "Resource Provisioning Policies to Increase IaaS Provider's Profit in a Federated Cloud Environment", High Performance Computing and Communications (HPCC), 2011 IEEE 13th International Conference.

[8] Hyun Jin Moon, Yun Chi, Hakan Hacigumus, "SLA-Aware Profit Optimization in Cloud Services via Resource Scheduling", IEEE 6th World Congress on Services, 2010.

[9] Linlin Wu , Saurabh Kumar Garg, RajkumarBuyya," SLA-based admission control for a Software-as-a-Service provider in Cloud computing environments", Journal of Computer and System Sciences vol. 78,pp.1280–1299,2012.

[10] K. Xiong, H. Perros, "SLA-based resource allocation in cluster computing systems, ": Proceedings of 17th IEEE International Symposium on Parallel andDistributed Processing (IPDPS 2008), Alaska, USA, 2008.

[11] B.N. Chun and D.E. Culler, "User-Centric Performance Analysis of Market-Based Cluster Batch Schedulers," Proc. Second IEEE/ ACM Int'l Symp.Cluster Computing and the Grid, 2002.

[12] Durkee, "Why Cloud Computing Will Never be Free," Comm. ACM, vol. 53, no. 5, pp. 62-69, 2010.

[13] R. Ghosh, K.S. Trivedi, V.K. Naik, and D.S. Kim, "End-to-End Performability Analysis for Infrastructure-as-a-Service Cloud: An Interacting Stochastic Models Approach," Proc. 16th IEEE Pacific Rim Int'l Symp. Dependable Computing, pp. 125-132, 2010.

[14] K. Hwang, G.C. Fox, and J.J. Dongarra, Distributed and Cloud Computing. Morgan Kaufmann, 2012.

[15] D.E. Irwin, L.E. Grit, and J.S. Chase, "Balancing Risk and Reward in a Market-Based Task Service," Proc. 13th IEEE Int'l Symp. High Performance Distributed Computing, pp. 160-169, 2004.

[16] Y.C. Lee, C. Wang, A.Y. Zomaya, and B.B. Zhou, "Profit-Driven Service Request Scheduling in Clouds," Proc. 10th IEEE/ACM Int'l Conf. Cluster, Cloud and Grid Computing, pp. 15-24, 2010.

[17] F.I. Popovici and J. Wilkes, "Profitable Services in an Uncertain World," Proc. ACM/IEEE Conf. Supercomputing, 2005.

[18] J. Sherwani, N. Ali, N. Lotia, Z. Hayat, and R. Buyya, "Libra: A Computational Economy-Based Job Scheduling System for Clusters, "Software - Practice and Experience, vol. 34, pp. 573-590, 2004.

[19] C.S. Yeo and R. Buyya, "A Taxonomy of Market-Based Resource Management Systems for Utility-Driven Cluster Computing," Software - Practice and Experience, vol. 36, pp. 1381-1419, 2006.

[20] A.P. Chandrakasan, S. Sheng, and R.W. Brodersen, "Low-Power CMOS Digital Design," IEEE J. Solid-State Circuits, vol. 27, no. 4, pp. 473-484, Apr. 1992.

[21] K. Li, "Optimal Load Distribution for Multiple Heterogeneous Blade Servers in a Cloud Computing Environment," Proc. 25th IEEE Int'l Parallel and Distributed Processing Symp. Workshops, pp. 943-952, May 2011.

[22] K. Li, "Optimal Configuration of a Multicore Server Processor for Managing the Power and Performance Tradeoff," J. Supercomputing, vol. 61, no. 1, pp. 189-214, 2012.

[23] P. Mell and T. Grance, "The NIST Definition of Cloud Computing," Nat'l Inst. of Standards and Technology, http://csrc.nist.gov/groups/SNS/cloud-computing/, 2009.

[24] B. Zhai, D. Blaauw, D. Sylvester, and K. Flautner, "Theoretical and Practical Limits of Dynamic Voltage Scaling," Proc. 41st Design Automation Conf., pp. 868-873, 2004.

[25] Yang, X. S.," Bat Algorithm for Multiobjective Optimization", Int. J. Bio-Inspired Computation, Vol.3, No.5, pp.267-274,2011.

[26] T. B¨ack, U. Hammel, and H.-P.Schwefel. Evolutionary computation: Comments on the history and current state. *IEEE Transactions on Evolutionary Computation*, 1(1):3–17, 1997.

[27] S. Bleuler, M. Brack, L. Thiele, and E. Zitzler.Multiobjective genetic programming:Reducing bloat by using SPEA2. In *Congress on Evolutionary Computation (CEC-2001)*, pages 536–543, Piscataway, NJ, 2001. IEEE.

[28] C. A. CoelloCoello, D. A. Van Veldhuizen, and G. B. Lamont.*Evolutionary Algorithms for Solving Multi-Objective Problems*. Kluwer, New York, 2002.

[29] D. W. Corne, J. D. Knowles, and M. J. Oates. The pareto envelope-based selection algorithm for multiobjective optimisation. In M. Schoenauer et al., editors, *ParallelProblem Solving from Nature (PPSN VI)*, pages 839–848, Berlin, 2000. Springer.

[30] M. P. Fourman. Compaction of symbolic layout using genetic algorithms. In J. J.Grefenstette, editor, *Proceedings of an International Conference on Genetic Algorithms and Their Applications*, pages 141–153, Pittsburgh, PA, 1985. Sponsored by Texas Instruments and U.S. Navy Center for Applied Research in Artificial Intelligence (NCARAI). Vol. 1993, pages 197–212, Berlin, 2001. Springer.

[31] S. Helbig and D. Pateva. On several concepts for _-efficiency.*OR Spektrum*,16(3):179–186, 1994.

[32] J. Horn, N. Nafpliotis, and D. E. Goldberg. A niched pareto genetic algorithm for multiobjective optimization. In *Proceedings of the First IEEE Conference on Evolutionary Computation, IEEEWorld Congress on Computational Computation*, volume 1, pages 82–87, Piscataway, NJ, 1994. IEEE Press.

[33] H. Ishibuchi and T. Murata.Multi-objective genetic local search algorithm. In *Proceedings of 1996 IEEE International Conference on Evolutionary Computation(ICEC'96)*, pages 119–124, Piscataway, NJ, 1996. IEEE Press.

[34] J. D. Knowles and D. W. Corne. The pareto archived evolution strategy: A new baseline algorithm for paretomultiobjective optimisation. In *Congress on Evolutionary Computation (CEC99)*, volume 1, pages 98–105, Piscataway, NJ, 1999.IEEE Press.

[35] F. Kursawe. A variant of evolution strategies for vector optimization.In H.-P.Schwefel and R. M¨anner, editors, *Parallel Problem Solving from Nature*, pages193–197, Berlin, 1991.Springer.

[36] G. Rudolph. *Convergence Properties of Evolutionary Algorithms*.VerlagDr.Kova˘c, Hamburg, 1997.

[37] G. Rudolph. Evolutionary search for minimal elements in partially ordered sets. In *Evolutionary Programming VII – Proc. Seventh Annual Conf. on Evolutionary Programming (EP-98)*, San Diego CA, 1998.The MIT Press, Cambridge MA.

Semantic Web Improved with the Weighted IDF Feature

Mrs. Jyoti Gautam
Department of Computer Science and Engineering
JSSATE (Uttar Pradesh Technical University)
NOIDA, U.P., INDIA

Dr. Ela Kumar
Department of Computer Science and Engineering
Indira Gandhi Delhi Technical University for Women
Delhi, INDIA

***Abstract*—The development of search engines is taking at a very fast rate. A lot of algorithms have been tried and tested. But, still the people are not getting precise results. Social networking sites are developing at tremendous rate and their growth has given birth to the new interesting problems. The social networking sites use semantic data to enhance the results. This provides us with a new perspective on how to improve the quality of information retrieval. As we are aware, many techniques of text classification are based on TFIDF algorithm. Term weighting has a significant role in classifying a text document. In this paper, firstly, we are extending the queries by "keyword+tags" instead of keywords only. In addition to this, secondly, we have developed a new ranking algorithm (JEKS algorithm) based on semantic tags from user feedback that uses CiteUlike data. The algorithm enhances the already existing semantic web by using the weighted IDF feature of the TFIDF algorithm. The suggested algorithm provides a better ranking than Google and can be viewed as a semantic web service in the domain of academics.**

Keywords—Text classification; Semantic Web with weighted idf feature; Expanded query; New Semantic Web Algorithm; Ranking Algorithm

I. INTRODUCTION

A lot of information is available on the Internet. Search engines remain as the primary infrastructure for Information Retrieval. The relevance of the result-sets is not as desired by the user. This leads to the requirement of a good ranking algorithm to put the best results on the front.

Many popular Web services like Delicious, Citeulike and flickr.com rely on folksonomies (Gautam and Kumar, 2012).Some websites such as CiteUlike (Research Paper Recommender), Delicious (online bookmarking), Flickr (online photo management and sharing application), Furl (File Uniform Resource Locators), Blinklist (links saver), Diigo (collect and organize anything e.g. bookmarks, highlights, notes, sceenshots etc.), Otavo (collaborative web search), Stumbleupon (discovery engine), Blummy (tool for quick access to favorite web services), and Folkd (saves bookmarks and links online) etc. which contain these tag information.

Various difficulties are encountered while doing research on folksonomies. In spite of all this, the growth is tremendous in this area. Researches based on social-bookmarking have become increasingly popular, which lets users specify their keywords of interest, or tags on web resources. Social tagging, also known as social annotation or collaborative tagging is one of the major characteristics of Web 2.0. Social-tagging systems allow users to annotate resources with free-form tags. The resources can be of any type, such as Web pages (e.g., delicious), videos (e.g., YouTube), photographs (e.g., Flickr), academic papers (e.g., CiteULIke), and so on.

In this paper, we utilize the semantic tag information with web page. This information is obtained from CiteUlike (Research Paper Recommender and online Tagging System).When users submit their query; they also submit some semantic description to disambiguate the query. Then, by matching the semantic description between the query and web page, user's query intent can be well understood. The better understanding of the user's query leads to better ranking results in academic domain.

In this paper, the following approach has been adopted. We have tried to use the metadata available in the form of user feedback and semantic tags from CiteUlike.

a) A new ranking algorithm has been developed. The algorithm utilizes the weighted IDF feature of the TFIDF algorithm.

b) The query was expanded. The idea was to use "keyword + tags" instead of keywords only, so that it carries some semantic description along with it.

c) The data was obtained through CiteUlike.

d) The performance analysis was done by comparing the approach with Google by several evaluation methods.

The paper is organized by an introduction to the existing ranking methods, then the new optimized JEKS algorithm followed by significance of the algorithm. Thereafter, the experiments and analysis is done followed by significance and relevance of the research work. In the end, finally the paper is concluded.

II. THE EXISTING RANKING METHODS

Tf-idf, term frequency-inverse document frequency is a numerical statistic which reflects how important a word is to a document in a corpus. The tf-idf value increases proportionally to the number of times a word appears in the document, but is offset by the frequency of the word in the corpus.

The literature (S. Lu, X. Li, S. Bai and S. Wang., 2000) provides an improved approach named tf.idf.IG to remedy this defect by Information Gain from Information Theory.

The literature (S. Lu, X. Li, S. Bai and S. Wang., 2000) provides an improved approach named tf.idf.IG to remedy this defect by Information Gain from Information Theory.

The Lingo algorithm proposed by Osinski and Weiss (2005) combines common phrase discovery and latent semantic indexing techniques to separate search results into meaningful groups. It looks for meaningful phrases to use as cluster labels and then assigns documents to the labels to form groups.

(Wu, Zhang and Yu, 2006) explored the technique of Social Annotations for the Semantic Web. These annotations are manually made by normal web users without a predefined formal ontology. The evaluation of the approach shows that the method can effectively discover semantically related web bookmarks that current social bookmark service cannot discover easily.

(Farooq, Kannampallil and Song, 2007) The authors use six tag metrics to understand the characteristics of a social bookmarking system. Possible design heuristics was suggested to implement a social bookmarking system for Cite Seer using the metrics.

The authors Cilibrasi and Vitanyi (2007) described a technique for calculating the Google similarity distance.

Jin, Lin and Lin (2008) proposed the architecture of a semantic search engine and an improved algorithm based on TFIDF algorithm. The algorithm considers crawling of static web pages. The algorithm can be considered for crawling of dynamic web pages and for parallel crawling also.

A personalized search framework was proposed by Shenliang, Shenghua and Fei (2008).It utilizes folksonomy for personalized search.

(Jiang, Hu, Li, and Wang 2009).The other method of basic TFIDF model uses supervised term weighting approach. The model uses class information to compute weighting of the terms. The approach is based on the assumption that low frequency terms are important, high frequency terms are unimportant, so it designs higher weights to the rare terms frequently.

Jomsri, Sanguansintukul and Choochaiwattana (2010) proposed a framework for Tag-Based Research Paper Recommender system. User self-defined tags were used for creating a profile for each individual user and cosine similarity was used to compare a user profile and research paper index. The recommender system demonstrated an encouraging preliminary result with the overall accuracy percentage up to 91.66%. The number of subjects is considered to be small in the experiment.

(Zhao and Zhang, 2010) proposed a new viewpoint on how to improve the quality of information retrieval. The queries are extended by "keywords+tags" instead of keywords only. A new tag based ranking algorithm (OSEARCH) was proposed and the results obtained were also compared with Google by several evaluation methods.

The authors Leung and Lee (2010) focussed on search engine personalization and developed several concept-based user profiling methods that are based on both positive and negative preferences. The proposed methods were evaluated against the previously proposed personalized query clustering method.

(Kaczmarek, 2010) introduced a novel approach to interactive query expansion. When a user executes a query, the algorithm shows potential directions in which the search can be continued.

Another supervised term weighting method, proposed by the authors (Zhanguo, Jing, Liang, Xiangyi and Yanqin, 2011), provides an improved tf-idf-ci model to compute weighting of the terms. The method uses intra and inner class information.

Various variations of the tf–idf weighting scheme are often used by search engines. Search engines use these weighted measures as a central tool in scoring and ranking a document's relevance given a user query. The tf-idf is improved by many literatures. The proportion of distribution of terms in text collection is one of the most important factors of expressing the content of text, but it is beyond tf-idf's power (Zhanguo, Jing, Liang, Xiangyi and Yanqin, 2011).

The paper proposed by (Yoo, 2011) suggests a hybrid query processing method for the effective retrieval of personalized information on the semantic web. When individual requirements change, the current method of query processing requires additional reasoning for knowledge to support personalization.

(Halpin and Lavrenko, 2011) proposed the method of relevance feedback between hypertext and semantic web search. The paper proposed investigates the possibility of using semantic web data to improve hypertext web search.

In this paper, the authors (Gracia and Mena, 2012) presented the web's natural semantic heterogeneity problems – namely, redundancy and ambiguity. The authors' ontology matching, clustering, and disambiguation techniques aim to bridge the gap between syntax and semantics for Semantic Web construction.

The authors Zhong, Li and Wu (2012) proposed an effective pattern discovery method for text mining. The paper presents an innovative and effective pattern discovery technique which includes the processes of pattern deploying and pattern evolving, to improve the effectiveness of using and updating discovered patterns for finding relevant and interesting information.

The paper (Lee, Kim and Park, 2012) proposes searching and ranking method of relevant resources by user intention on the semantic web. There are more limitations in information searching as the information on the Internet dramatically increases. To overcome the various limitations, the Semantic Web must provide search methods based on the different relationships between resources.

This paper proposed by (Gautam and Kumar, 2012) proposes a framework for a tag-based Academic Information Sharing and Recommender System which shares information such as question papers, assignments, tutorials and quizzes on a specific area.

(Shaikh, Siddiqui and Shahzadi, 2012) proposed the Semantic Web based Intelligent Search Engine. SWISE required including domain knowledge in the web pages to answer intelligent queries. The layered model of Semantic Web provides solution to this problem by providing tools and technologies to enable machine readable semantics in current web contents.

(Lee, Kim, and Park 2012) presented some proposals to improve and extend the semantic approach based on conceptual neighborhood's graphs in order to best preserve the proximity between the adapted and original documents and to deal with models that define delays and distances.

III. User Query Intent and Storage of Tags

A. Metadata Information in the Web Pages and Expansion of the Query

While talking about semantic web, metadata comes into picture. What is this semantic? How is it related to metadata? Semantic Web is something that implies the content, meaning or the metadata related to the web. This metadata information is hidden in the web pages. There are different websites which are working upon it since a long time. We have sites like Delicious, CiteUlike, Flickr etc., which allow different users to create their accounts. After creating the accounts, the users can add metadata for the different websites. This metadata conveys the content of the website as interpreted by different users.

The method should be such that which tries to capture the user's real query intent. The primary purpose of the search engines is to return the optimal results. But before returning the results, it should be able to analyze the query clearly. The simple keywords can't express user's real query intent. In order to analyze the query, some metadata information is added along with the query. The metadata information is added by expanding the query .i.e., keyword+tags instead of the keywords only.

So, the idea is to consider utilizing metadata which is available in the form of semantic tags .One area that arises is to consider utilizing the semantic tag information with web page. When users submit their query, they can also submit some simple semantic description to narrow down the query. Then by matching the semantic information between query and web page metadata, we can understand user's query intent better and return better result.

So, the idea is to utilize this semantic tag information. Here, we are proposing the development of a new algorithm based on semantic tags and the weighted IDF feature of the TFIDF algorithm.

B. Storage of Semantic Tags on Web Pages

The semantic tags of a web page are some object properties that reflect the content of the web page, such as marked with "semantic web", which signifies that the page contains information about the object of "semantic web". Of course, there may be multiple tags on a page, because the pages always contain multi information. These tags carry the metadata information along with them.

In our case, we are storing the tags from CiteUlike. A popular website in academia is CiteULike (www.CiteULike.org). CiteUlike is a free service for managing and discovering scholarly references.

- Easily store references you find online
- Discover new articles and resources
- Automated article recommendations
- Share references with your peers
- Find out who's reading what you are reading
- Store and search your PDF's

CiteULike has a filing system based on tags. Tags provide an open, quick and user-defined classification model that can produce interesting new categorizations.

Additionally, it is also capable to:

- 'tag' papers into categories.
- Add your own comments on papers.
- Allow others to see your library

The semantic tags are retrieved from CiteUlike. The URLs along with their tags are stored in a local database. For the semantic tags, each URL is opened in CiteUlike and the tags with their numeric values are stored in the database. We add tags' values in the MYSQL database. The data was retrieved from April, 2012 to June, 2013 from CiteUlike for the 50 queries. A total of 5000 URLs were opened in CiteUlike and the database was created.

IV. A new Optimized Ranking Algorithm

A. Utilizing the Weighted Inverse Document Frequency

In this paper, we are proposing a new algorithm based on semantic tags in the web pages. An enhanced semantic web algorithm is proposed. The algorithm is based on utilizing the metadata information available with the web pages by integrating in the algorithm some good features of weighted IDF.

Here, we are improving the semantic web by utilizing the weighted IDF score. We are already familiar with (1), which is applicable in the context of TFIDF (Jiang, Hu, Li, and Wang, 2009)

$$W(tk,dj,ci) = (1-\alpha).tfidfk,j + \alpha. tfidfk,j \times weighting \quad (1)$$

$$weighting = Ai/Ci, \quad \text{(Refer TABLE 1.)} \quad (2)$$

α is called a balance factor, which lies between , $0 \leq \alpha \leq 1$.

When $\alpha = 0$, (1) becomes classic TFIDF approach, and when $\alpha = 1$, (1) becomes our newly improved approach. Using balance factor, we can get better classification results.

TABLE I. Below shows the Relation of Term T_K and Category C_i.

	C_i	$\bar{C}_i$
t_k	A	B
$\bar{t}_k$	C	D

A indicates the number of documents belonging to category C_i where the term t_k occurs at least once; B indicates the number of documents not belonging to category C_i where the term t_k occurs at least once; C denotes the number of documents belonging to category C_i where the term t_k does not occur at least once; D denotes the number of documents not belonging to category C_i where the term t_k does not occur at least once.

This equation (1) is applicable for the terms of the document. The same equation can be used for tags also. Let us take an example. For the three tags, tag1, tag2, tag3 of the category Ci, if they share the same values of tf-idf but have different proportion of A and C. So, the tags which have higher values of the weighting factor make more contribution to the category Ci. Evidently, the tf-idf approach gives equal weights to the three tags unlike the weighted ones.

Now, we have integrated this equation with the other equation proposed by Zhao and Zhang (2010)

B. A New Optimized Ranking Algorithm – JEKS (Jyoti and Ela Kumar Search) algorithm

Initially, when users want to submit a query, instead of just giving the query in the form of keywords, they will also expand the query by adding some metadata information along with the query. Afterwards, the algorithm compares the inputted tags in query with the semantic information on the web pages in order to provide the user with better results.

Accordingly, the user query can be expressed as:

Query = {keyword1, keyword2,…, tag1, tag2,…}

In the above formulation, keyword1, keyword2 is the main query keyword.Tag1; tag2 is the semantic information which we are adding to expand the query. For example, Query = {research papers, web mining) represents that the user wants to find information relating to research papers on web mining.

Similarly, Query = {resources, information retrieval}

represents that the user wants to find information relating to resources in the field of information retrieval.

Once, the query is submitted, the system creates a vector of all the user tags.

V_usrt = {user_tag1, user_tag2,…}

Once the query is submitted to the search engine, the engine returns an initial result page list. The vector of all the tags on the result pages is recorded.

V_rest = {r_tag1, r_tag2,…}

Where, r_tag1, r_tag2 represent semantic tags on result pages.

The similarity is calculated between the two tag vectors, and recorded as a Tg_score.

Then, the final score of the web page is:

TotalScore=google_score+(Tg_score*IDFscore*weighting) (3)

Score=Tg_score*IDFscore*weighting (4)

Re – rank the google results according to this score.

Here, google_score represents the original google results score when the query is applied.

Google_score=(p-q+1)/p (5)

Here, p represents the total no. of documents, which is 100 in the experiment; q represents the location of the document on search engine's result list. So, google_score for the 6th result is (100- 6 + 1) / 100 = 0.95.

In (3), Tg_score is calculated by matching the tags of the user with the tags of the result page. The match between the two vectors is based on the following factors.

1) The similarity between the user tag vector and web page tag vector. The high value is obtained by high similarity between the two vectors.

2) The other factor being the weight of the tags on the result pages. Weight refers to the frequency of the the tags in the result pages which match with the tags of the user.

Tg_score is defined as given below based on the factors considered:

$$\text{Tg\_score}=\frac{\sum_{i=1}^{|\text{V\_usrt}|}\sum_{k=1}^{|\text{V\_rest}|}(freq(\text{V\_rest[i]})*\text{sim}(\text{V\_usrt[i]}),\text{V\_rest[k]}))}{\sum_{k=1}^{|\text{V\_rest}|}freq(\text{V\_rest[k]})} \quad (6)$$

In the above equation, freq (tag) represents the frequency or weight of the particular tag on the result page. $\text{sim}(\text{V\_usrt[i]}),\text{V\_rest[k]})$ represents the similarity between the user tag vector V_usrt[i] and the result page tag vector V_rest[k] and similarity is defined as given below:

$$\text{sim}(\text{V\_usrt[i]}),\text{V\_rest[k]})$$

= 1, V_usrt[i] and V_rest[k] have the same root,
= 1, V_usrt[i] and V_rest[k] have the same meaning,
= 0, V_usrt[i] and V_rest[k] does not have a semantic relation,
=0.5, even if half of the V_usrt[i]tag resembles with the V_rest[k]tag. (7)

,e.g. let us say in the Query = {resources, information retrieval} , resources is the keyword and information retrieval is the tag, then in the tags of the result pages even if information or retrieval appears , we have taken the similarity score as 0.5.

Next, ,e.g. consider the query , Query = {artificial intelligence, pdf} to Google, The tenth result has the tags as "pdf", "pdfs", "research" and the frequency of the tags is 10, 9, 4 respectively. Then, the value of the Tg_score = (10*1+9*1+4*0)/ (10+9+4) = 19/23 and google_score = (100-10+1)/100=0. 91.

Next in (3) is the IDF score multiplied by weighting. We know from the TFIDF algorithm.

Given a document collection D, a word w, and an individual document d Є D, we calculate

$$w_{d=}f_{w,d}*\log(|D|/f_{w,D}), \quad (8)$$

Where $f_{w,d}$ equals the number of times w appears in d, |D| is the size of the corpus, and $f_{w,D}$ equals the number of documents in which w appears in D. Words with high w_d imply that w is an important word in d but not common in D.

Here, if the above equation is analyzed properly, we see that if we replace words with tags, the (8) can be used in the context of semantic web. So, $f_{w,d}$ has already been considered

as the Tg_score. Now remains the log ($|D|/f_{w,D}$), (which is IDF score). Here, for each query, we have taken the 100 Google results. So, for a particular query, D is 100 and $f_{w,D}$ equals the number of documents in which the particular tag of the query appears.

Now, why we have included this IDF score?

Suppose that Tg_score is large and fw,D score is small. Then log ($|D|/f_{w,D)}$ will be rather large, and so in (3), the score will be large. This is the case we are most interested in, since tags with high score imply that this tag is important for the document d but not common in D. This tag is having a large discriminatory power. Therefore, when a query contains this tag, returning a document d where score is large will very likely satisfy the user.

Now, we are multiplying this IDF score with the weighting factor. As, we have already mentioned the significance of this weighting factor .Let us take an example. Let us replace the terms with the tags in (8).If the values of (Tg_score * IDF score) is similar for the different tags, then weighting factor is used to differentiate the results. The tags with the higher weighting will be preferred as they have higher discriminating power for the category Ci in comparison to the tags having less weighting factor. The tags having less weighting may be rare tags in the category Ci.

Now, calculating the (IDF score * weighting factor) for the Query = {books, artificial intelligence}, let us say that the documents in which the tag artificial intelligence appears is 30 and the value of D is 100. So, the IDF score is log (100/30) and weighting factor is (30/70).

In the above (6), we are using java functions to calculate the similarity between user tags and result tags. The database is created using MYSQL.

For example, user submits the query "research papers, mobile computing", to Google, the 4th result of Google is having the tag's values, mobile computing = 37, mobile devices = 35, mobile interaction = 27, pedestrian navigation = 23, navigation = 12. And, the tag mobile computing appears in 37 documents. So, according to the above algorithm, the total score = (0.97) + (0.507) * log (100/37)*(37/63).

V. Significance of the JEKS Algorithm

The JEKS algorithm developed above is effective in the case when (Tg_score*IDF) score is similar for the different tags in a category Ci. Through the values of the proportion of Ai and Ci, it can be easily found that the three tags show different discriminating power to TC (Refer TABLE 2). The weighting factor can be used to differentiate the results. The tags with the higher weighting will be preferred as they have higher discriminating power for the category Ci in comparison to the tags having less weighting factor. The tags having less weighting may be rare tags in the category Ci. For example, take a class Ci as research papers and the three different tags as mobile computing, data mining and semantic web. Corresponding to this, the three different queries are {research papers, mobile computing}, {research papers, data mining} and {research papers, semantic web}. Now for a particular case when Tg_Score and IDF Score is similar for the three different tags of the class Ci, then Ai/Ci will be used to produce three different TotalScore values (Refer (3))., and hence different rankings.

TABLE II. THREE TAGS WHICH SHARE THE SAME (TG_SCORE*IDF SCORE) BUT HAVE DIFFERENT PROPORTION OF AI AND CI IN A CATEGORY CI

Tag	Tg_Score	IDF Score	Ai/Ci
Tag1(mobile computing)	.507	Log(100/37)	2:10
Tag2(data mining)	.507	Log(100/37)	1:1
Tag3(semantic web)	.507	Log(100/37)	10:2

The TABLE 2 shows that the tag3 gives higher discriminating power to the category Ci from other categories than the tags tag1 and tag2. The tag1 may be a rare tag in the category Ci, and makes little contribution to the category Ci. So, the TotalScore will be highest for the tag3, lowest for tag1 and for tag2, it lies in between.

VI. Experiments and Analysis

The experiments are performed as follows:

1) Initially, submit the query to Google, and obtain the original Google search results.

2) Now, submit the Google search results to CiteUlike to obtain the relevant tags.

3) Re-rank the search results according to our algorithm.

4) Compare the Google results with our algorithm.

A. Data Set

Query Set: Initially, we determine the queries which we input to the search engine. We determine a total of fifty queries. The queries are a combination of keywords and tags. These queries are submitted to Google. The queries are from academic domain as CiteUlike provides tags for the academic database.

Result Set: Now, submit each query to Google and record the first 100 results. This way, the result set of 50 queries become 5000 results.

Results Tag Set: Now, we submit the 5000 results to CiteUlike and the resulting tag vector is recorded. We obtain lots of tag values for a result.

For example, user submits the query "resources, genetic algorithm", to Google, the 4th result of Google is having the tag's values, genetic algorithm = 37, genetic = 35, algorithm = 27, pedestrian navigation = 23, navigation = 12. And, the tag genetic algorithm appears in 40 urls. So, according to the above algorithm, the total score = (0.97) + (0.507) * log (100/40)*(40/60).

We have chosen the following queries.

Q1 = {books, artificial intelligence}
Q2 = {books, grid computing}
Q3 = {books, information retrieval}
Q4 = {books, java programming}
Q5 = {books, software engineering}
Q6 = {pdf, artificial intelligence}
Q7 = {pdf, cloud computing}
Q8 = {pdf, data structure}
Q9 = {pdf, deep web}
Q10 = {pdf, digital image processing}
Q11 = {pdf, distributed computing}
Q12 = {pdf, parallel algorithm}
Q13 = {pdf, semantic web}
Q14 = {research papers, communication}
Q15 = {research papers, compiler}
Q16 = {research papers, data mining}
Q17 = {research papers, genetic algorithm}
Q18 = {research papers, mobile computing}
Q19 = {research papers, pharmacology}
Q20 = {research papers, quantum cryptography}
Q21 = {research papers, semantic web}
Q22 = {research papers, software engineering}
Q23 = {research papers, statistics}
Q24 = {research papers, ubiquitous computing}
Q25 = {research papers, web mining}
Q26 = {research papers, wireless communication}
Q27 = {resources, electronics engineering}
Q28 = {resources, grid computing}
Q29 = {resources, information retrieval}
Q30 = {resources, semantic web}
Q31 = {resources, ubiquitous computing}
Q32 = {books, automata}
Q33 = {books, data mining}
Q34 = {books, power electronics}

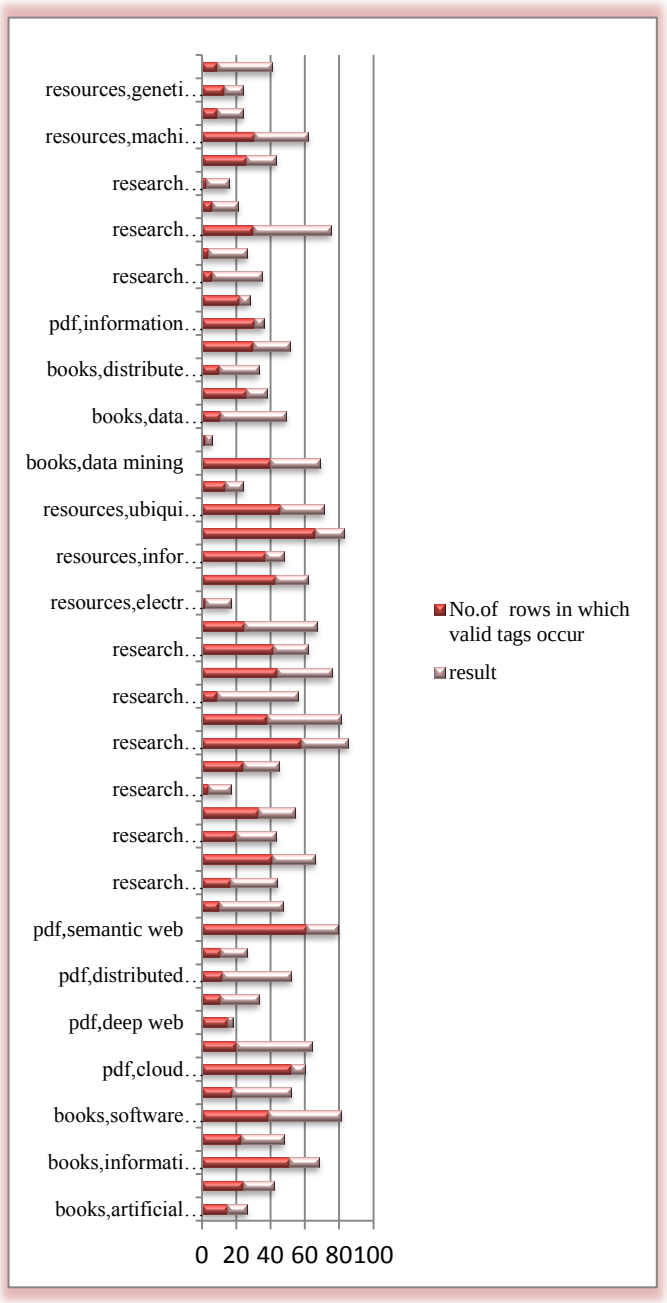


Fig. 1. The number distribution of specific tags versus difference tags in a result set

Rows in which valid (specific) tags occur = A
Rows of total tags = B, Difference Tags (result) = C = (B-A)

e.g., for the query = {research papers, data mining} A = 41, B = 66, C = 25.

Q35 = {books, data structure}
Q36 = {books, deep web}
Q37 = {books, distributed computing}
Q38 = {books, web mining}
Q39 = {pdf, information retrieval}
Q40 = {pdf, genetic algorithm}
Q41 = {research papers, digital signal}
Q42 = {research papers, fluid mechanics}
Q43 = {research papers, machine learning}
Q44 = {research papers, molecular electronics}
Q45 = {research papers, power electronics}
Q46 = {resources, database}
Q47 = {resource, machine learning}
Q48 = {resources, molecular electronics}
Q49 = {resources, genetic algorithm}
Q50 = {resources, structure analysis}

B. Experimental Results

First, we determine the relevance between each query intent and each result page. Each result is assigned a relevance score according to its relevance, which ranges between 0 to 3 (totally irrelevant, basically irrelevant, basically relevant, and totally relevant).

We obtain normalized DCG values for our algorithm and Google as given in the Table 3.

TABLE III. COMPARISION OF NORMALIZED DCG (*NDCG*) VALUES FOR OUR ALGORITHM AND GOOGLE

QUERY NO.	nDCG(A)	nDCG(G)
q1	0.957424	0.970031
q2	0.888747	0.913824
q3	0.877744	0.862172
q4	0.938299	0.934294
q5	0.854472	0.881374
q6	0.887192	0.885906
q7	0.975138	0.97113
q8	0.86662	
q9	0.834386	0.796038
q10	0.920252	0.942012
q11	0.959862	0.953069
q12	0.995585	0.995332
q13	0.982126	0.981987
q14	0.897661	0.84126
q15	0.881929	0.848669
q16	0.933084	0.894468
q17	0.975616	0.983474
q18	0.908892	0.85308
q19	0.805438	0.801738
q20	0.929742	0.91508
q21	0.945845	0.938982
q22	0.92802	0.913109
q23	0.879856	0.770643
q24	0.956999	0.945143
q25	0.83687	0.760944
q26	0.934957	0.92141
q27	0.928905	0.928905
q28	0.994868	0.994253
q29	0.957072	0.964861
q30	0.993879	0.992997
q31	0.986664	0.984467
q32	0.877298	0.837017
q33	0.911324	0.905407
q34	0.934142	0.934407
q35	0.91458	0.902485
q36	0.900831	0.940498
q37	0.87344	0.941972
q38	0.887338	0.854466
q39	0.983348	0.976529
q40	0.982363	0.981797
q41	0.907483	0.851692
q42	0.840634	0.790858
q43	0.905059	0.883763
q44	0.856174	0.818381
q45	0.969632	0.969562
q46	0.961921	0.943682
q47	0.986109	0.979038
q48	0.986892	0.983111
q49	0.98629	0.981852
q50	0.855997	0.882663

We obtained normalized DCG values for the 50 queries for our algorithm as well as for Google results. We observed that Fig. 2 shows the normalized DCG values of 50 queries. The graph compares our algorithm with Google. It can be seen that our algorithm acquires higher values of DCG for 40 queries when compared to Google.

Next, we use Precision@k curve for various Relevance levels.

The following conclusion can be drawn from the Fig. 3 to Fig. 5.Our algorithm acquires higher precision in comparison

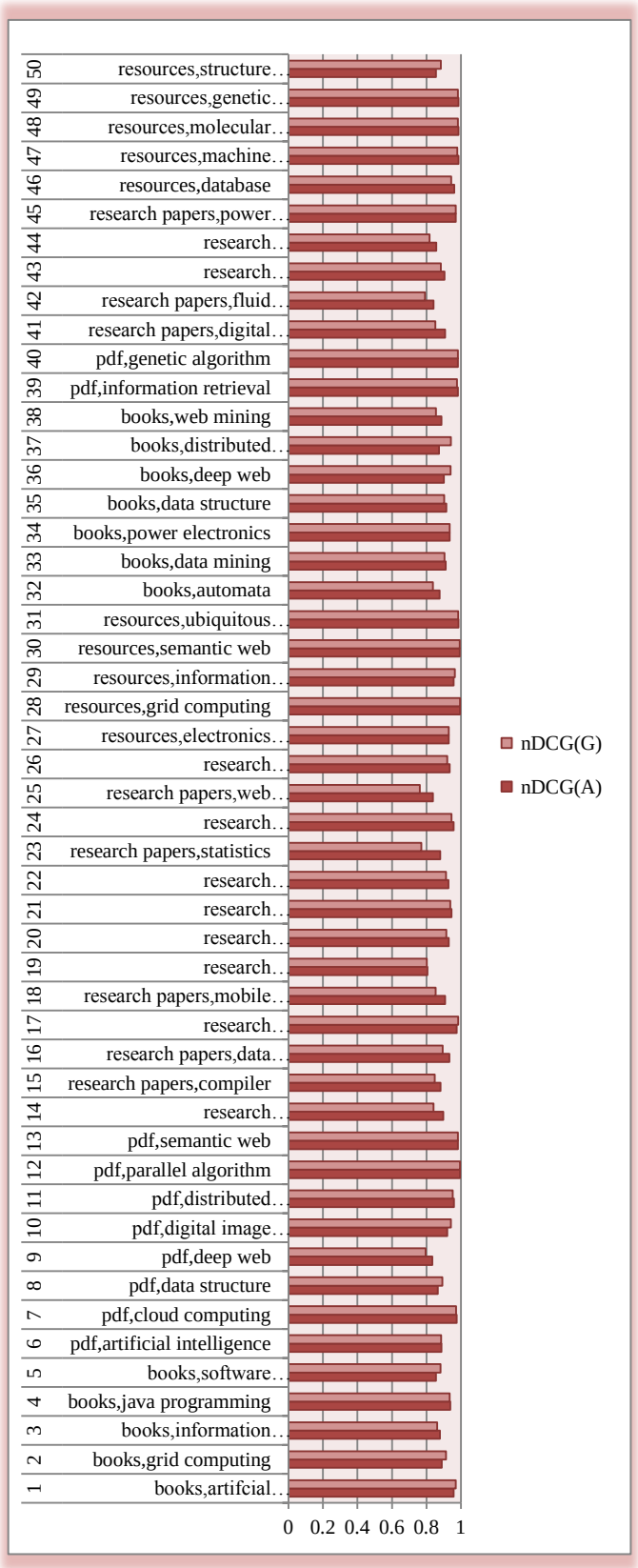


Fig. 2. The average DCG value of 50 queries

to Google throughout the varying levels of K for all the 50 queries. The results obtained for Rel>=1 are the best as expected. The precision for Rel>=1 are better than Rel>=2, which is better than Rel>=3.Only, when the Rel>=3, initially Google results are better as can be seen from Fig. 5.

We computed the values for precision, recall and F1-score for our algorithm and Google (Table 4.). These values are calculated for all the queries. These values are calculated for their corresponding top 50 results for Rel>=2 for all the 50 queries. We observed that the value of recall for our algorithm and Google remain at 1 as we have re-ranked the top 100 results of Google for each query. The value of precision and F1-score are calculated and it has been observed that we are getting better results.

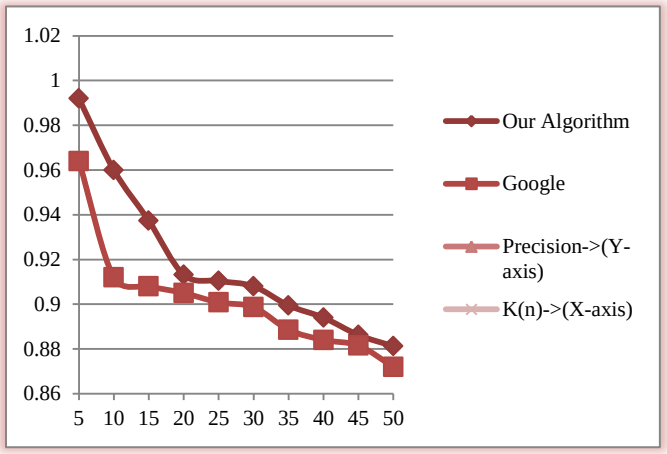


Fig. 3. The Precision@k curve of 50 queries when Rel>=1

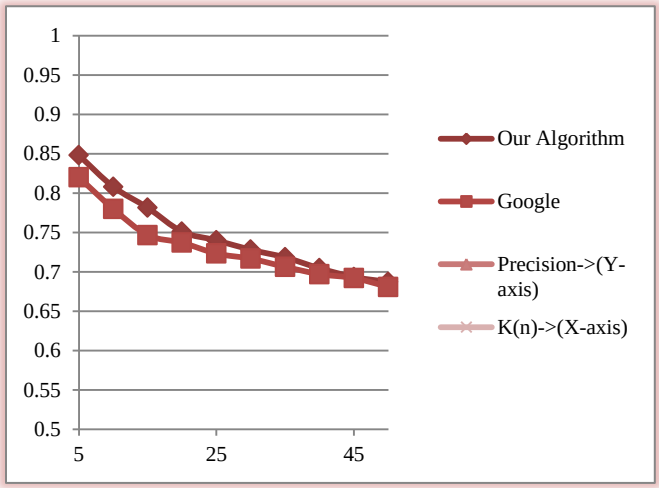


Fig. 4. The Precision@k curve of 50 queries when Rel>=2

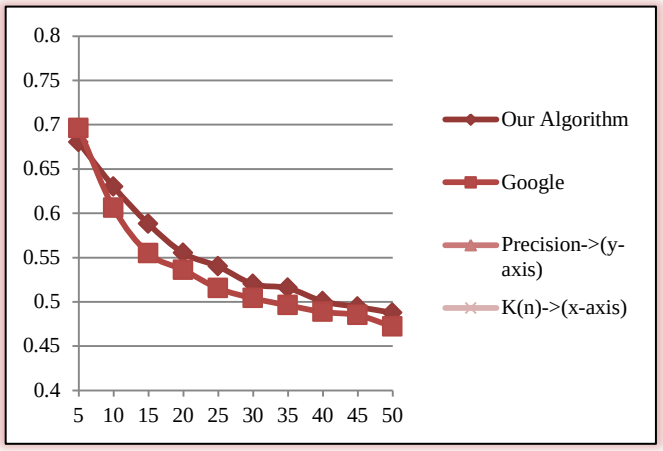


Fig. 5. The Precision@k curve of 50 queries when Rel>=3

TABLE IV. PRECISION AND F1-SCORE FOR OUR ALGORITHM AND GOOGLE

	JEKS algo		Google	
Query	*PRECISION*	*F1-score*	*PRECISION*	*F1-score*
q1	0.94	0.969	0.96	0.98
q2	0.5	0.667	0.5	0.667
q3	0.34	0.507	0.38	0.551
q4	0.72	0.837	0.72	0.837
q5	0.72	0.837	0.7	0.824
q6	0.8	0.889	0.8	0.889
q7	0.96	0.98	0.94	0.969
q8	0.5	0.667	0.5	0.667
q9	0.32	0.485	0.3	0.462
q10	0.9	0.947	0.9	0.947
q11	0.88	0.936	0.86	0.925
q12	0.96	0.98	0.96	0.98
q13	0.98	0.99	0.98	0.99
q14	0.56	0.718	0.56	0.718
q15	0.5	0.667	0.48	0.649
q16	0.64	0.78	0.62	0.765
q17	0.92	0.958	0.88	0.936
q18	0.72	0.837	0.72	0.837
q19	0.26	0.413	0.24	0.387
q20	0.7	0.824	0.68	0.81
q21	0.86	0.925	0.86	0.925
q22	0.66	0.795	0.62	0.765
q23	0.42	0.592	0.42	0.592
q24	0.8	0.889	0.78	0.876
q25	0.48	0.649	0.48	0.649
q26	0.74	0.851	0.68	0.81
q27	0.6	0.75	0.6	0.75
q28	1	1	1	1
q29	0.84	0.913	0.84	0.913
q30	1	1	1	1
q31	0.94	0.97	0.94	0.97
q32	0.54	0.701	0.5	0.667
q33	0.6	0.75	0.62	0.765
q34	0.42	0.592	0.42	0.592
q35	0.16	0.276	0.18	0.305
q36	0.42	0.592	0.42	0.592
q37	0.72	0.837	0.72	0.837
q38	0.46	0.63	0.5	0.667
q39	0.96	0.98	0.96	0.98
q40	0.96	0.98	0.96	0.98
q41	0.44	0.611	0.42	0.592
q42	0.42	0.592	0.42	0.592
q43	0.6	0.75	0.54	0.701
q44	0.52	0.684	0.54	0.701
q45	0.92	0.958	0.92	0.958
q46	0.8	0.889	0.76	0.864
q47	0.92	0.958	0.92	0.958
q48	0.98	0.99	0.98	0.99
q49	0.94	0.969	0.94	0.969
q50	0.42	0.592	0.42	0.592

VII. SIGNIFICANCE OF THE RESEARCH WORK

Being an academician, I preferred to work in the Academic Domain. I have selected some 50 queries applicable in the Academic Domain. The queries are focused on retrieving the books in different fields of computer science, research papers in different fields of electronics and computers, resources in the respective fields and pdf in various fields of computers. I have retrieved Google results for those queries. For a single query, I have retrieved first 100 results. Those 100 urls were submitted to CiteUlike for retrieving metadata (i.e. tags). In totality, I have retrieved 5000 urls and the tags corresponding to those urls with their weights. The Google results were re-ranked corresponding to those queries using my algorithm.

After this, I had applied JEKS algorithm on 5000 urls(corresponding to 50 queries). My results of JEKS algorithm for normalized DCG for 40 queries (out of 50 queries) were higher than Google. Our algorithm acquires higher precision in comparison to Google throughout the varying levels of K for all the 50 queries.

We computed the values for precision, recall and F1-score for our algorithm and Google .These values are calculated for all the queries. These values are calculated for their corresponding top 50 results for Rel>=2 for all the 50 queries.

We observed that the value of recall for our algorithm and Google remain at 1 as we have re-ranked the top 100 results of Google for each query. The value of precision and F1-score are calculated and it has been observed that we are getting better results.

So, the significance of my research work is that a better ranking system has been developed using my algorithm for retrieving the results in academic domain. The results can be extended to include more queries.

VIII. Relevance of My Research Work

The relevance of the research work is that the entire work has been done using semantic tags from CiteULike(which provides tags in a fully uncontrolled environment). The algorithm is entirely based on tags, which are the essence of semantic web. So, it can be taken as an application or a web service in Academics Domain using semantic web. The algorithm can be extended for more queries.

IX. Conclusion

In this paper, we have analyzed some existing ranking methods and proposed a new algorithm based on the previous methods. Semantic tag of a web page is the metadata information associated with it and depicts a lot about the information associated with it. The match degree between user's real query intent and web page content is determined by calculating the similarity between query and web page tag.

We have proposed the new algorithm using the already existing semantic web algorithm which basically calculates the weighted score of the tags. We have utilized the IDF feature of TFIDF algorithm to improve the semantic web which uses tags. In addition to this, we have used a weighting score. In experiments, we have collected the data from Citeulike and implemented the above algorithm. The relevance scores to the different web links have been given by a group of users. Comparing with Google search results, we find that JEKS algorithm acquires better ranking results, and can put more relevant results in front. Our algorithm acquires higher values of DCG for 40 queries when compared to Google. Our algorithm acquires higher precision in comparison to Google throughout the varying levels of K for all the 50 queries.

In the future work, we will further improve the algorithm. We will consider combining with the search engines user logs, and mining out information repeated to user's query, such as the click information, the browse information and so on. The algorithm can be further enhanced by adding these effects. The algorithm can be extended to include more queries.

References

[1] Cilibrasi, R.L., & Vitanyi, P.M.B. (2007) The Google similarity distance, in IEEE Transactions on Knowlegde and Data Engineering, vol. 19, no.3.

[2] Farooq, U., Kannampallil, T.G., & Song, Y. (2007) Evaluating Tagging Behaviour in Social Bookmarking Systems: Metrics and design heuristics, *in the international ACM Conference on Supporting Group Work*.

[3] Gautam, J., & Kumar, E. (2012) An Improved Framework for Tag-Based Academic Information Sharing and Recommender System, in *Proc. of the World Congress on Engineering*, Vol. 2, 2012, 845-850.

[4] Gracia, J., & Mena, E. (2012) Semantic Heterogeneity Issues on the Web, *IEEE Internet Computing*, pages 60-67.

[5] Halpin, H., & Lavrenko, V. (2011) Relevance feedback between hypertext and Semantic Web search, *Journal of Web Semantics*, vol. 9, 2011, pages 474-489.

[6] Jiang, H., Hu, X., Li, P., & Wang S.(2009) An improved method of term weighting for text classification, in *International Conference on Intelligent Computing and Intelligent Systems*, IEEE, Vol.1, 2009, pages 294-298.

[7] Jin Y., Lin Z., & Lin H., The Research of Search Engine Based on Semantic Web, in *proc. of International Symposium on Intelligent Information Technology Application Workshops (IITAW), IEEE*, 2008, pages 360-363.

[8] Jomsri P., Sanguansintukul S. & Choochaiwattana W., A Framework for Tag-Based Research Paper Recommender System: An IR Approach, in *proc. of the 24th International Conference on Advanced Networking and Applications Workshops, IEEE*, 2010, pages 103-108.

[9] Kaczmarek, A.L. (2011) Interactive Query Expansion with the Use of Clustering-by-Directions Algorithm, *IEEE Transactions on Industrial Electronics*, VOL. 58, No. 8, pages 3168-3173.

[10] Lee, M., Kim, W., & Park, S. (2012) Searching and ranking method of relevant resources by user intention on the Semantic Web, Expert Systems with Applications, vol. 39, pages 4111- 4121.

[11] Leung, K.W.T., & Lee, D.L. (2010) Deriving concept-based user profiles from search engine logs, in *IEEE Transactions on Knowledge and Data Engineering*, vol. 22, no. 7.

[12] Lu C., Hu X., & Park J. (2011) Exploiting the Social Tagging Network for Web Clustering, *(Systems, Man, and Cybernetics – Part A: Systems and Humans)*, vol. 41, pp. 840-852.

[13] Maredj A., & Tonkin N. (2013) Semantic Adaptation of Multimedia-Documents, International Arab Journal of Information Technology, vol. 10, No. 6, pages 579-586.

[14] Osinski, S., & Weiss, D. (2005) A Concept-Driven Algorithm for Clustering Search Results, *IEEE Intelligent Systems*, Volume 20, Issue 3, pp. 48-54.

[15] Shaikh, F., Siddiqui, U.A. & Shahzadi, I. (2012) Semantic Web based Intelligent Search Engine, in *proc. of International Conference on Information and Emerging Technologies*, pp. 1-5.

[16] S. Lu, X. Li, S. Bai & S. Wang., (2000) An improved approach to weighting terms in text. *Journal of Chinese Information Processing*, 14(6), pp. 8-13.

[17] Shenliang X., Shenghua B. and Fei, B., *Exploring Folksonomy for Personalized Search*, in proc. of the 31st annual international ACM SIGIR conference on Research and Development in information retrieval, 2008, pp. 155-162.

[18] Yoo, D. (2012) Hybrid Query Processing for Personalized Information Retrieval on the Semantic Web, *Knowledge-Based Systems*, vol 27, pages 211-218.

[19] Wu, X., Zhang, L., & Yu Y. (2006) Exploring Social Annotations for the Semantic Web, in *proc. of the 15th International Conference on World Wide Web (WWW 06), ACM*, pages 417-426.

[20] Zhanguo, M., Jing, F., Liang, C., Xiangyi H., & Yanqin, S. (2011) An improved approach to terms weighting in text classification , in *proc. of the International Conference on Computer and Management*, IEEE, pages1-4.

[21] Zhao, C., & Zhang, Z. (2010) A New Keywords Method to Improve Web Search, in *12th International Conference on High Performance Computing and Communications, IEEE*, pages 477-484.

[22] Zhong, N., Li, Y., & Wu. S.T. (2012) Effective Pattern Discovery for Text Mining, IEEE Transactions on Knowledge and Data Engineering, vol. 24, no1.

Consuming Web Services on Android Mobile Platform for Finding Parking Lots

Isak Shabani, Besmir Sejdiu, Fatushe Jasharaj
Department of Computer Engineering
Faculty of Electrical and Computer Engineering
University of Prishtina
Prishtina, Republic of Kosovo

***Abstract*—Many web applications over the last decade are built using Web services based on Simple Object Access Protocol (SOAP), because these Web services are the best choice for web applications and mobile applications in general. Researches and the results of them show how architectures and the systems primarily designed for use on desktop such as Web services calls with SOAP messaging, now are possible to be used on mobile platforms such as Android.**

The purpose of this paper is the study of Android mobile platform, more precisely the ability of this platform for consuming Web services and exploring existing alternatives for consuming Web services from this platform. People use their vehicles every day for transport and this of course leads to a constant demand for finding a parking lot. In this paper is proposed the system, named as *MyParking* through which it is aimed to facilitate users finding a parking lot for their vehicle depending on their current location. *MyParking* consists of three modules: Android client, administration and Web services.

Keywords—Web application; Web services; Android platform; Mobile devices; MyParking

I. INTRODUCTION

Smart mobile devices are increasingly popular the last few years and the total number of mobile devices sold globally is constantly growing. At the same time, the computing power of these devices is growing at an incredible pace and for several years has already reached the power of desktop personal computers. Since many of the limitations of previous generations of mobile devices such as limited memory and persistent storage capacity, limited CPU power, as well as limited and intermittent internet connection and bandwidth are not prevalent anymore, these devices nowadays can be used for advanced applications [1].

Along with the popularity of mobile devices and their growth are created a range of platforms and applications programming environments for them. Different platforms of operating systems like Symbian OS, PalmOS, J2ME, Blackberry, Windows Mobile, iOS and Android, are currently used by equipment vendors in their mobile devices. All these platforms require specific programming language or dialect specific to the implementation of applications. Basically, the only option that is independent from the platform is Java ME for virtual machines which exist in most recent platforms. Unfortunately Java ME is quite old and most of the existing restrictions, are used in its design. This makes Java ME somewhat outdated today but it is the only technology-independent platform. Android supports a set of Java APIs, it uses Java as programming language, it has a broad support of adaption, it has built in components for the graphic of user's interface and a set of key applications available built from third-party developers. A range of service platforms were recently built based on Service Oriented Architecture with SOAP messaging protocol. By creating mobile clients for these platforms it is possible a greater support of system using [2].

Since majority of commercial applications are not pure mobile applications but they rather use mobile clients in distributed and complex software systems, there exists an emergency need for at least implementation of the independent server by distributed application platforms. Web services often are used to provide such an implementation. As majority of clients mobile applications require more subtle interactions and in this way they are often implemented as Representational State Transfer (RESTful) Web services, semantically rich interfaces of commercial applications often use Web services based in SOAP. Hence, in order to be able to use mobile devices of recent time in business processes, it`s important the support of SOAP Web services in platforms of mobile devices. The support for SOAP Web services is not perfect in platforms mentioned above nowadays but there exists other ways to use such services [1]. Whenever an application offers a type of interface that can be called programmatically from another application by sending commands through the Hyper Text Transfer Protocol (HTTP), is said to be an example of Web services [3].

Applications in mobile devices need to communicate with others system components by consuming Web services. Therefore the aim of this project is to explain how is possible to be consumed these services in Android platform, respectively the consumption of SOAP Web services.

II. BACKGROUND AND RELATED WORK

One of the usual functions that is required in mobile applications is to call a Web service to draws the data. This process includes searching of Web service with parameters of getting response. There are two different types of Web services: SOAP and RESTful.

- *SOAP services* usually have a defined contract that is signed or is followed with all structures of data, service

methods and not only them. This contract is written in Web Services Description Language (WSDL) and is published for costumers that use Web services. Also these types of services mainly use EXtensible Markup Language (XML) for requirements and data response.

- *RESTful services* are more ad-hoc than SOAP services because they don't use WSDL and they are based on the standards preliminarily established as XML and HTTP. These types of services are free to restore the data in every possible format, and the communication between them is easier.

A. Consuming RESTful Web services on Android

RESTful web services are simple, scalable, easy to use, attuned to the philosophy of the Web, and able to handle a wide variety of clients [4]. In technical level, Web services can be implemented in Android. Before the implementation particular client of Web services should take in consideration that mobile devices are limited by bandwidth of network and power which is based on the battery. There a lot of headers and layers of SOAP elements in the XML load. So the usage of SOAP services unlike the usage of RESTful client Web services in Android devices is more costly, as for the developer, so for the user. Further, Android SDK offers support for consumption of RESTful Web services by offering libraries/packages in form of HTTP client.

B. Consuming SOAP Web services on Android

There is a considerable number of Web services based on SOAP that are consumed by mobile applications. Especially in the world of enterprises, applications in mobile devices need to communicate with components of other systems by consuming Web services. Android doesn't offer native support for consumption of Web services, but exist a useful library called kSOAP2 which permits Android applications that in an easy and efficient way to consume Web services based on SOAP [5]. This library is third-party library distributed as free source, optimized for Android [6].

In the proposed system MyParking, consume of Web services in Android is realized through kSOAP2. In fact, kSOAP2 [7] is only a project that simplifies usage of SOAP in Android. This library encompasses details of basic layer of transport, offers different mechanisms for (de) serialization of different messages and facilitates handling of SOAP defects. Libraries should be added in project in order to be used. This library is based in SOAP architecture and there is no need to generate any proxy/stub to call Web service methods [6].

There exist different applications for finding parking lots in different countries.

S. Srikanth et.al, [9] proposed a Smart Parking (SPARK) Management System which provides advanced features like remote parking monitoring, automated guidance & parking reservation mechanism. Though prototype system, they proposed the architecture which satisfies the car parking management system requirement.

S. Khang et.al, [10] proposed a parking system in which driver comes to know about the space availability in the parking lot with the help of SMS service. Driver can resend SMS in order to request new space if the previous one is filled. Driver can find nearest space for parking using wireless mobile based car parking system. Results, shows that the system efficiently allocates the slots and utilizes the full parking space.

G. Yan et.al, [11] proposed NOTICE based parking system. In this parking system, drivers can check and reserve the slot for parking. For security purposes encryption/decryption techniques are used. Simulation results are highly efficient.

In Kosovo, actually doesn't exist any application, in any mobile platform for finding parking lots, therefore the proposed system in this paper will find a wide application.

III. System Design

MyParking is an Android application that helps users to find parking lots depending on their location. The main purpose of this application is to offer to users facilities to use application which helps them to find parking places depending on their location. Except MyParking module for clients, there exists also administration module and Web services for communication between client module and the server as well as Web services for communication between parking lots and the server. In the context of this paper client application is an application that is executed in Android mobile platform and which accesses the SOAP Web services server.

A. System Architecture

System consists from administration module and client's module. Administration module is developed in Microsoft .NET Framework 4.0/ASP.NET platform and as programming language is used *C#*. For developing client application on Android platform is used *Java* programming language and other components which are needed to develop Android applications such as *Eclipse* with *ADT plugin* and *Android SDK*. For exchanging the data between client and server are used *SOAP Web services* developed in *ASP.NET 4.0* platform.

System architecture as is shown in Fig. 1 consists of three main parts: Android client, the server and parking lots. In Android client is made registration of client that uses the application, search of parking by proximity, city and address and also the visualization of the data. In server is made managing of parking lots, cities and clients, where from parking lot are sent parking details such as number of free places, prices, the total number of places and identification code of parking lot.

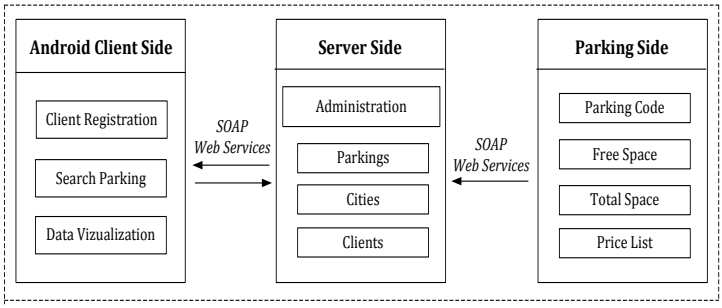


Fig. 1. System architecture

B. Network architecture

In Fig. 2 is presented system network architecture that consists of database server, the server in which is published Web application of administration module, the server in which are published Web services that are used from parkings for sending free places and the server in which are published Web services that are consumed from Android devices (clients) for data visualization about parking places that consist: location, number of free places, the total number of places, price and address.

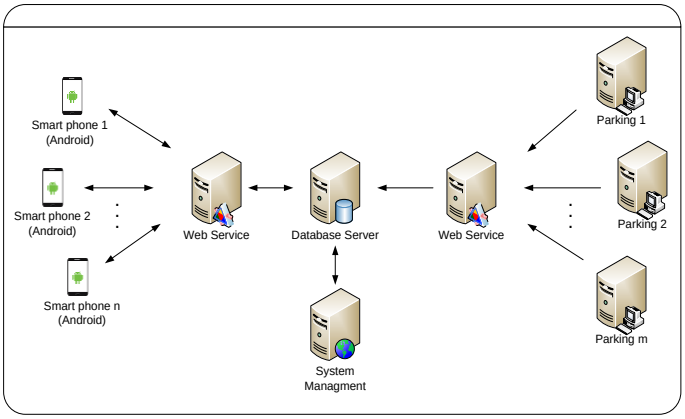


Fig. 2. System network architecture

IV. Database Design

System contains server database which is in MS SQL Server platform and the client database SQLite. In the Table I is presented the list of table's databases of the server and their description.

TABLE I. Description of the Database Server Tables

#	Table Name	Description
1	Parking	Contains information of parking places
2	Cites	Contains the list of the cities
3	Clients	Contains the clients which use the application
4	ClientVisits	Contains the client visits
5	Administration Users	Contains the list of administration module users
6	SyncData	The table which contains information about data syncronization between the client and the serve

Table II presents the list of the database tables of the client and their description, whereas the diagram of database is presented in Fig. 3.

TABLE II. Description of the Client Database Tables

#	Table Name	Description
1	Parking	Contains informations about parking places which are obtained from server
2	Cities	Containsthe list of the cities which are obtained from server
3	ClientInfo	Contains informations of the client that use the application in his device.
4	SyncData	Table which contains informaions about syncronization of database between the client and the server

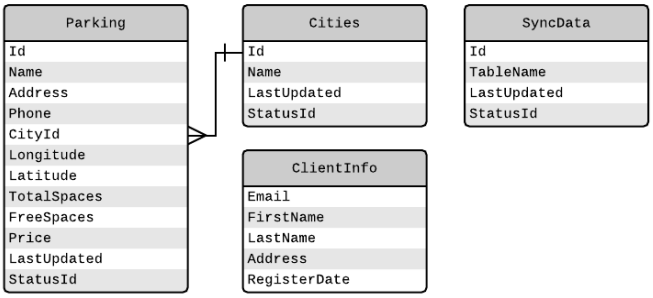


Fig. 3. Diagram of the server database - SQLite

V. Data Syncronization Through Web Services

Communication between *Parkings* and the *Server* is done through *SOAP Web services* [8], the flow of the data is only in one direction, Parking - Server. These Web services consist of two web methods:

- PostParkingFreeSpace
- PostParkingData

PostParkingFreeSpace – consists of two parameters: *parkingCode* (parking code that identifies which parking is sending data) and *freeSpace* (number of free places that actually are in the parking). This web method is called by parkings every time when in parking changes the number of free places.

PostParkingData – consists of three parameters: *parkingCode* (parking code that identifies, which parking is sending data). *totalSpace* (total number of places which are in parking). This Web method is called rarely from parkings, only in cases when the parking changes the total number of places or changes the prices.

Communication between *Android client* and the *Server* is also made by SOAP Web services *Android client* sends and receives data from the server.

Two web methods are developed for sending data in server:

- PostClient
- PostClientVisit

PostClient – through this web method client's data are sent in server in the case of registration. The sent data are: e-mail, first name, last name and registration date.

PostClientVisit – web method is called from Android client when the user opens the Android application, to send in the server the information that the application is used. The data which are sent are: e-mail, date, Android version of client's device and also his actual geographical position (latitude and longitude).

To take the data from the server are developed three web methods:

- ListSyncTables
- ListCities

- ListParkings

Web method *ListSyncTables* lists tables for synchronization, such as the name of the table, and the date of the last update (insert/update) of that table:

TABLE III. LIST OF THE DATA THAT WEB METHOD LISTSYNCTABLES RESTORES

Table Name	Last Updated
Cities	2014-09-30 22:32:50.963
Parkings	2014-10-02 14:00:58.850

Based on this list is easy to understand that which of the tables had changes of the data by comparing the last update with the date which is in SyncTables table in SQLite (in Android device). If the last update which is in the SQLite table is older than the date of the list which restores *ListSyncTables* web method for the specific table, for example Parkings, then is understood that in that table were some changes of the data and the synchronization of the data in that table is needed.

Web method *List Cities* – as input parameter accepts the date which is taken from the table of client synchronizations (where the name of the table is "*Cities*") which can be found in SQLite and restores the list of cities from the server - only in those cities where are added/modified after this date of input parameter. After wards these cities are added/modified in the table *Cities* of the client database SQLite. Synchronization of the cities is accomplished in this way. Similarly functions also parking synchronization by using *ListParkings* Web method.

VI. THE INTERFACE

A. Client application MyParking

Client application MyParking is designed by user's view. User friendly design helps users to achieve their aims. Efforts and aim has been that design must be very simple and understandable for the users. Client applications forms are designed in XML and business logic is written in Java. Google Maps API is used to make easier for the user to find the parkings in the nearest distance with his current position. Users won't need to try a lot to understand the functionality and navigation in the application. Following are presented forms and main characteristics of client application:

- Registration Form
- Home page
- The search of parking lots by proximity
- The search of parking lots by city
- The search of parking lots by address
- The parking lot details form

1) Registration form

Registration form appears to the user only once, after application is installed. Through this form user is registered by giving data such as: first name, last name and e-mail. These data are automatically sent to server through Web services. The Fig. 4 shows the registration form.

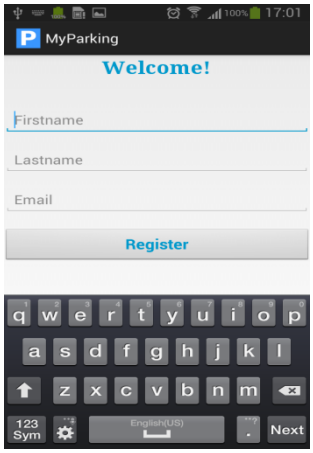


Fig. 4. Registration form

2) Home page

The Home page form is displayed after user`s registration. If the user is registered earlier, this form opens as application`s starting form. At the top of this form is displayed user`s current location on the map, and five nearest parkings if any, while at the bottom are buttons for advanced search (search by proximity, city and address) and the button to update data. This form is shown in Fig. 5.

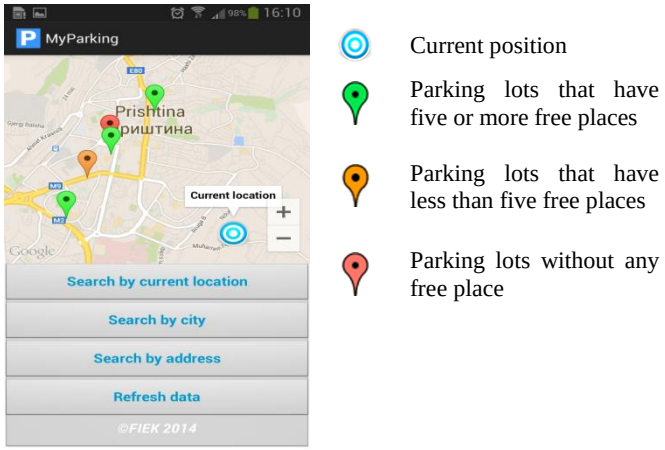
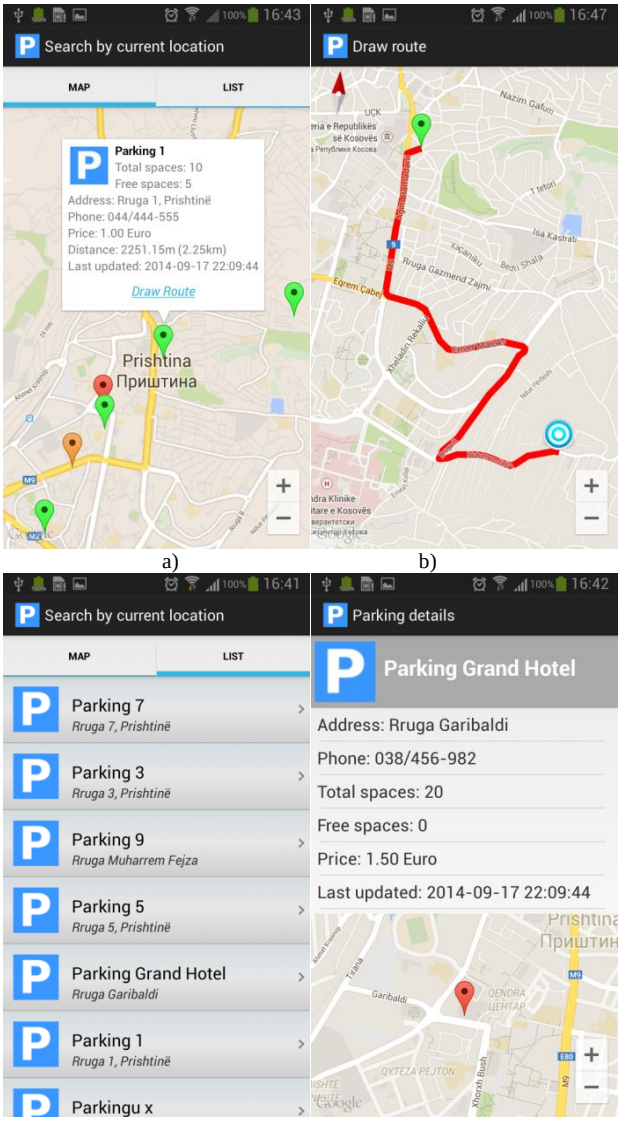


Fig. 5. User's Home page

3) The search of parking lots by proximity

Search for parking lots by proximity is enabled by clicking 'Search by proximity' button from *Home page*. Through this form user is able to see nearest parking lots in visual form in the map or in the list form. Fig. 6 a) shows parking lots close to the current position. By clicking on the particular parking lot, appears the window that contains information about that parking as well as 'Show the path' link through which opens the form as shown in Fig. 6 b) which shows the path form current position to the selected parking lot. The search for parking lots by proximity appears also in the list form as shown in Fig. 6 c), where parking lots are displayed sorted according to proximity. By clicking on particular parking lot opens the parking lot details form as in Fig. 6 d).



a) b)

c) d)

Fig. 6. Search according to proximity: a) parking lots map, b) the path from current position to the selected parking lot, c) a list of parking lots, d) details of parking lot

4) The search of parking lots by city

In this form are listed all cities of Kosovo. To see all parking lots of one city the certain city is selected and the form that contains parking lots opens in visual way on the map and the list form. The Fig. 7 on the left shows the search of parking lots by cities, while Fig. 7 to the right shows on the map parking lots of selected city, which also can be displayed on the list form, similar to the Fig. 6c and 6d.

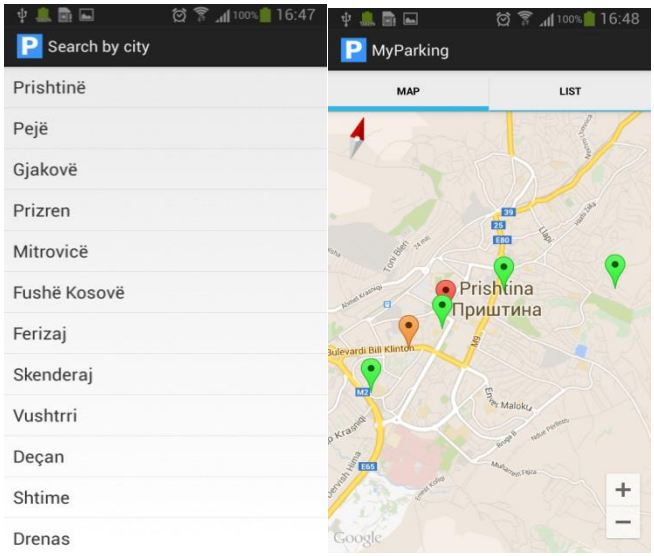


Fig. 7. Search by city form

5) The search of parking lots by address

Through this form the user is allowed to search for parking lots by address. The Fig. 8 shows the form of such a search.

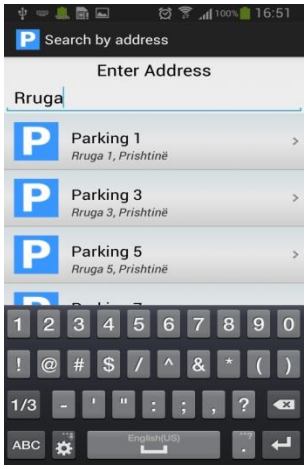


Fig. 8. Search by address form

6) The parking lot details form

Once the parking lot from any of the forms above is selected, than details of parking lot such as: parking lot name, the address, city, phone, total number of parking places, the number of free spaces and the date (timestamp) when the information about parking lot are lastly taken (update date) are displayed. At the bottom of this form appears the parking lot on the map. The Fig. 9 shows the details of parking lot form.

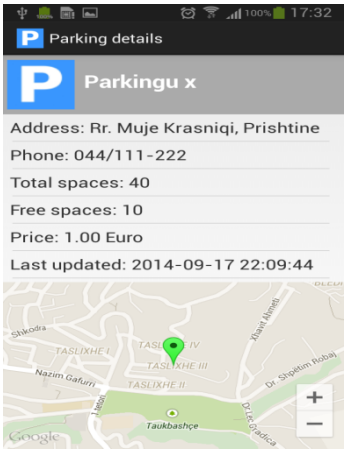


Fig. 9. The form of parking lot details

B. *The administration module*

Management (recording/editing) of parking lots and cities is possible through this module. Through this module also is possible the displaying of customers that use client application.

To login in the administration module is necessary that in advance, the user (administrator) to log in the system by giving username and the password. After the user is logged on the system, a form as in the Fig. 10 opens in which appear parking lots and their details, on the entire territory of Kosovo.

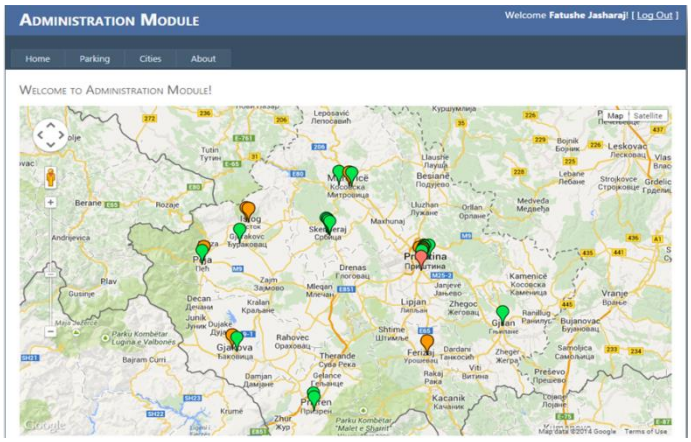


Fig. 10. The map of parking lots in the administration module

Via the form in Fig. 11, administrator registers parking lots and their relevant details, in which case a 5 digit code is generated that identifies the parking lot. When the parking lot sends data on the server about the number of free spaces, also sends this code to identify himself.

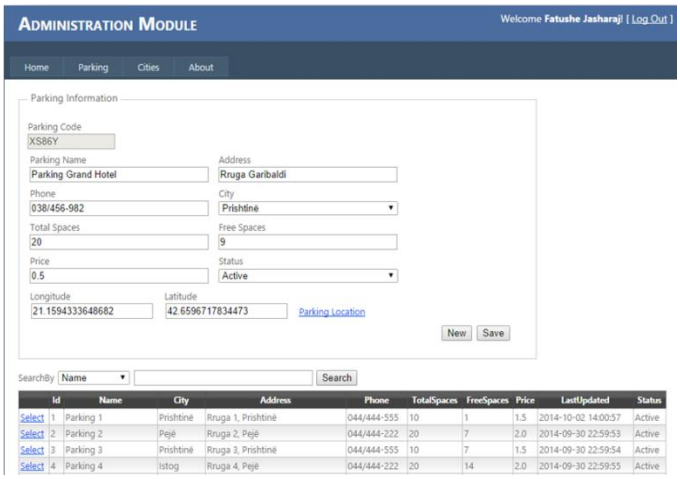


Fig. 11. The form of parking lots management

VII. Conclusion and Future Work

The processing capabilities of mobile devices have increased enormously in the recent years. This makes it possible to build more complex applications that target mobile devices. Recent searches and studies show how architectures and systems primarily designed for use on desktop such as Web services calls with SOAP messaging, now are available to be used on mobile platforms such as Android. With the help of faster and available mobile networks, direct access on Web services with SOAP messaging is definitely possible on Android.

Achieved results in this paper are successful adaptations of the solution, respectively of kSOAP2 library by client side for Web services on Android. It is also important to mention that kSOAP2 works fine when it`s imported into Android, despite the fact that some doubts are raised from developers about the necessity of rewriting some classes of APIs.

Based on the obtained results from our simulation study, we conclude that as from the server side as well as the client side, data synchronization is quick, optimal and effective.

Also, from the obtained results, we conclude that the proposed system can alleviate traffic congestion and reduce the amount of traffic volume caused by searching for parking.

In the future we will try to work and also explore other possible alternatives for consuming Web services on the Android platform and their support for this platform and other mobile platforms.

References

[1] C. Kleiner and Th. Schneider, “Securing SOAP Web Services for Mobile Devices on Different Platforms,” MMS 2011: Mobile und ubiquitäre Informationssysteme. Proceedings der 6. Konferenz, 28. Februar 2011 in Kaiserslautern, Deutschland. GI 2011 LNI ISBN 978-3-88579-279-6, vol. 185, pp. 25-38, 2011.

[2] J. Knutsen, “Web Service Clients on Mobile Android Devices, A Study on Architectural Alternatives and Client Performance,” p. 19, June 2009.

[3] U. Cei and P. Lucidi, “Alfresco 3 Web Services Build Alfresco applications using Web Services, WebScripts, and CMIS”, Packt Publishing, ISBN 978-1-849511-52-0, p. 8, 2010.

[4] L.Richardson and S. Ruby, “RESTful Web Services”, Published by O’Reilly Media, Inc., ISBN: 0-596-52925-0, p. 299-314, 2007.

[5] Th. Weerasinghe, Introduction to working with kSOAP2 in Android, thiranjith.com. [Accessible on October 20, 2014]

[6] P. Pocatilu, "Developing Mobile Learning Applications for Android using Web Services, Informatica Economica Journal, Vol. 14, No. 3/2010, pp. 106-115, 2010.

[7] J. Bertram and C. Kleiner, "Secure Web Service Clients on Mobile Devices", MobiWIS 2012: The 9th International Conference on Mobile Web Information Systems, vol. 10, pp. 696-704, 2012.

[8] I. Shabani, B. Çiço and A. Dika, "Solving Problems in Software Applications through Data Synchronization in Case of Absence of the Network" IJCSI International Journal of Computer Science Issues, Vol. 9, Issue 1, No 3, January 2012.

[9] S. Srikanth, P. Pramod, K. Dileep, S. Tapas, U. Patil, N. Babu, "Design & Implementation of a Prototype Smart Parking (SPARK) System using Wireless Sensor Networks", International Conference on Advanced Information Networking & Applications workshops, 978-0-7695-3639-2/09, IEE, 2009.

[10] S. Khang, T. Hong, T. Chin and Sh. Wang, "Wireless Mobile-Based Shopping Mall Car Parking System (WMCPS)", Services Computing Conference (APSCC), IEEE Asia-Pacific, pp. 573-577, 2010.

[11] G. Yan, S. Olariu, M.C. Weigle and M. Abuelela, "SmartParking: A Secure and Intelligent Parking System Using NOTICE", Intelligent Transportation Systems, ITSC 2008, 11th International IEEE Conference on, pp. 569-574, 2008.

[12]

Improvement of Control System Performance by Modification of Time Delay

Salem Alkhalaf
Computer Department, College of Science and Arts
Qassim University
Alrass City, Qassim, KSA

***Abstract*—This paper presents a mathematical approach for improving the performance of a control system by modifying the time delay at certain operating conditions. This approach converts a continuous time loop into a discrete time loop. The formula derived is applied successfully to an applicable control system. The results show that the proposed approach efficiently improves the control system performance. The relation between the sampling time and the time delay is obtained. Two different operating conditions are examined to assess the proposed approach in improving the performance of the control system.**

Keywords—Distributed control system; control delay; sampling scheme; control system performance

I. INTRODUCTION

A distributed control system (DCS) has many interconnected devices, which exchange data through the communication networks, such as home automation, factories, space shuttles, and industrial control Ethernet. The performance of the DCS network is assessed according to the ability of the network and its communication link to transmit the signal (bits) through the network with minimum delay and distortion. A delay in the process should be considered when designing the DCS network. The control loop performance over a network control system has been investigated and studied [1][2][3][4]. The real-time system, control system, and communication system have been studied. The performance of the control loop DCS network depends on many factors such as communication protocols, reducing the communication with dead bands, sampling time, and scan time. The time delay is considered when a design methodology for optimizing the performance of the distributed control system is presented [5]. Compensation for delay time uncertainties on industrial control Ethernet network has been investigated to prove how important delay time is in the control system performance [6]. The control system delay is the summation of the sensor to controller delay, controller calculation delay, and controller to actuator delay.

An adaptive sampling scheme has been presented [7] to ensure that the control delay is less than the sampling period in the steady state and uses the maximum tolerable delay at a specified sampling period to ensure stable transformation from one sampling interval to another.

Echo state neural networks have been used to improve the shape recognition performance of the sendzimir mill control system [8]. Modification of the control system based on artificial intelligence has been used successfully to improve the performance of existing coal-fired thermal power plants. A parameter prediction model based on an artificial neural network has been used to analyze the effect of advanced control on the combustion process, which lead to the development of a self-learning controller [9].

An adaptive robust control law for linear systems with norm-hounded parameter uncertainties has been developed for a robust control system. Online information of the actual system is used to tune the interpolation coefficient. This control scheme overcomes the shortage in a linear robust control with fixed parameters [10].

The neuro-fuzzy approach has been applied to the delay compensator to reduce variable sampling to actuation delays effect in the distributed control system. The approach proposed adding a compensator to an existing distributed system to overcome the degradation of the control performance that results from the variable sampling to actuation delay [11].

An improvement of existing coal fired thermal power plants performance by control systems modifications is discussed. The such system is applied via implementation of advanced combustion control concepts in selected Western Balkan thermal power plant, and particularly those based on artificial intelligence as part of primary measures for nitrogen oxide reduction in order to optimize combustion and to increase plant efficiency. Advanced self-learning controller has been developed and the effects of advanced control concept on combustion process have been analyzed utilizing artificial neural-network based parameter prediction model [12]**.** A discrete-time control systems performance has been optimized based on network-induced delay. The technique solving optimal tracking problem for single-input single-output (SISO) linear time-invariant discrete-time systems over communication channel with network-induced delay in the feedback path [13]**.** The output feedback control problem of an interconnected time-delay systems with prescribed performance has been solved and investigated. To obtain such valid solution, a few of the existing results consider the prescribed performance control in the nonlinear interconnected time-delay systems [14].

Keeping the sensor data validity while exercising timely control is crucial in real-time sensing and control systems. The objective of scheduling algorithms deployed in such systems is to keep the validity of the real –time sensor data. This approach leads to maximize the schedule ability of update transactions with minimum update workload, hence the control system

performance is adapted to be active on time [15]. Robust iterative learning control design for uncertain time-delay systems based on a performance index has been investigated, analyzed and discussed [16]. A robust iterative learning control (ILC) for uncertain time-delay systems has been designed based on a performance index for the error system. The Lyapunov-like approach can be applied to design robust ILC for uncertain systems with time-varying delay or multiple time delays. Enhancing the performance bounds of the multivariable control systems have high degree of thinness have been discussed, analyzed, and investigated [17]. A real-time implementation of fault-tolerant control (FTC) systems with performance optimization have been discussed, investigated and analyzed [18].

This paper aims to present a mathematical approach for improving the performance of a control system by converting the system continuous loop into a discrete loop. The formula derived is applied successfully to the studied system. The results show that the sampling time and the delay time are optimized. The sampling time against the time delay is an interested point.

II. Control System Model

The control system is shown in Figure 1 [19]. The simplest model is given by discrete-time control models obtained from continuous time models that include a constant time delay in the mathematical formulation. The continuous time state space model of the linear time invariant system can be described by the following standard form [20][21]:

$$dx(t)/dt = Ax(t) + Bu(t) \quad (1)$$

$$y(t) = cx(t) + Du(t) \quad (2)$$

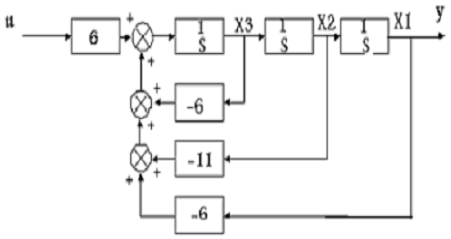


Fig. 1. Control System Block Diagram

Applying equations 1 and 2 to the control system, the system equations can be obtained in the matrix form:

$$\begin{bmatrix} X'1 \\ X'2 \\ X'3 \end{bmatrix} = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ -6 & -11 & -6 \end{bmatrix} \begin{bmatrix} X1 \\ X2 \\ X3 \end{bmatrix} + \begin{bmatrix} 0 \\ 0 \\ 6 \end{bmatrix} u \quad (3)$$

And the output equation is:

$$y = [1 \quad 0 \quad 0] \begin{bmatrix} x1 \\ x2 \\ x3 \end{bmatrix} \quad (4)$$

III. Prepare Delay Time and Sampling Time Approach

The classical model for discrete-time control systems assumes that the control algorithm is executed instantaneously at every sampling period h. Consequently, equidistant sampling and actuation are assumed.

Based on these assumptions, for periodic sampling with constant period h, the discrete time system can be described by:

$$x(kh+h) = x\,kh\,e^{Ah} + \frac{B}{A}[e^{A(h-\tau)} - 1]u\,kh + \frac{B}{A}e^{Ah}(1 - e^{-A\tau})u(kh - h) \quad (5)$$

Where:

h: Sampling time,

k: Number of control loop execution,

τ: Delay time,

A: Feedback transmission factor.

Multiplying equation 5 by B/A gives:

$$x\frac{A}{B}[kh + h - kh\,e^{Ah}] + ukh + e^{Ah}uh = e^{-A\tau}e^{Ah}hu \quad (6)$$

Reducing the previous equation gives:

$$e^{-A\tau} = \frac{A}{B}e^{-Ah}\frac{x}{u}(k - ke^{Ah} + 1) + k\,e^{-Ah} + 1 \quad (7)$$

Assume:

$$\frac{x}{u} = y$$

$$e^{-A\tau} = \frac{A}{B}e^{-Ah}(k - ke^{Ah} + 1)\,y + k\,e^{-Ah} + 1 \quad (8)$$

$$\tau = -\frac{1}{A}\ln[\frac{A}{B}e^{-Ah}(k - ke^{Ah} + 1)y + ke^{-Ah} + 1] \quad (9)$$

A is always –ve (negative feedback)

Thus, the equation above is valid if and only if

$$\frac{A}{B}e^{-Ah}(k - ke^{Ah} + 1)y + ke^{-ah} > 0 \quad (10)$$

or

$$\frac{A}{B}(k - ke^{Ah} + 1)y + k > 0 \quad (11)$$

i.e.,

$$y < \frac{B}{A}\frac{-k}{(1+k-ke^{Ah})} \quad (12)$$

IV. Simulation Results

Case I

From the loop of Figure 1 and the substitution in equation 12 by the given constant A=–6, B=6, k=32, h=0.001……0.005 sec, the results are shown in Table 1.

TABLE I. Sampling Time Against Delay Time

h	y	τ
0.001	26	0.118
0.002	23	0.034
0.003	20	0.078
0.004	18	0.05
0.005	16	0.107

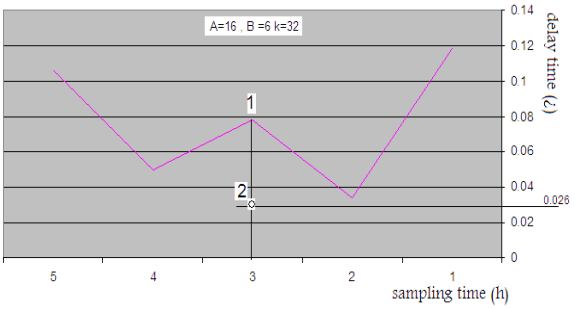


Fig. 2. A=–6 Primary Rough Relation Before Correction

Figure 2 depicts the primary rough relation between the control system delay time and the sampling time. Point 1 is irregular so piece wising of the curve will give a new point as (0,003,0.026). The table after correction is shown in Figure 2.

TABLE II. Sampling Time Against Delay Time After Correction

h	y	τ
0.001	26	0.118
0.002	23	0.034
0.003	20	0.026
0.004	18	0.050
0.005	16	0.108

The least square root method is used for the results after correction, which are given in Table 2.

Related to the quadratic equation, y=a+bx+cx2, where a, b, and c are constants, normal equations are:

$$N\,a + b\sum h_i + c\sum h_i^2 = \sum \tau_i \quad (13)$$

N=number of points (h's).

Substituting the values of h, n, and τ in equation 13, a, b, and c are obtained:

a=0.23, b=–136, c=22571

These values are substituted in equation 13.

$$\tau = 0.23 - 136\,h + 2257\,h^2 \quad (14)$$

Differentiate the previous equation:

$$\frac{d\tau}{dh} = 0 = -136 + 45142h$$

h=0.003, τ =0.025

Thus, the lowest point is (0.003,0.025).

Taking the different values of h and substituting them in equation 14 give the results shown in Table 3.

TABLE III. Sampling Time Against Delay Time

h	y	τ
0.001	26	0.118
0.002	23	0.034
0.003	20	0.026
0.004	18	0.050
0.005	16	0.108

h	τ
0.001	0.117
0.002	0.048
0.003	0.025
0.005	0.114

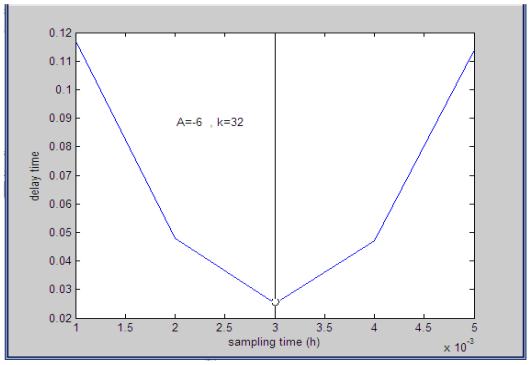


Fig. 3. Final Sampling Time Against Delay Time (A=–6, k=32)

Figure 3 depicts the primary rough relation between the control system delay time and the sampling time after correction. Comparing Figure 2 and Figure 3 shows the efficiency of the proposed approach in improving the control system performance to obtain an optimized point.

Case 2

The second operating point is considered A=–11, B =6, k=32, h=0.001,0.002,…,0.005 sec. The results are shown in Table 4.

TABLE IV. Sampling Time Against the Delay Time

h	y	τ
0.001	12	0.109
0.002	10	0.059
0.003	8.5	0.019
0.004	7	0.085
0.005	6	0.108

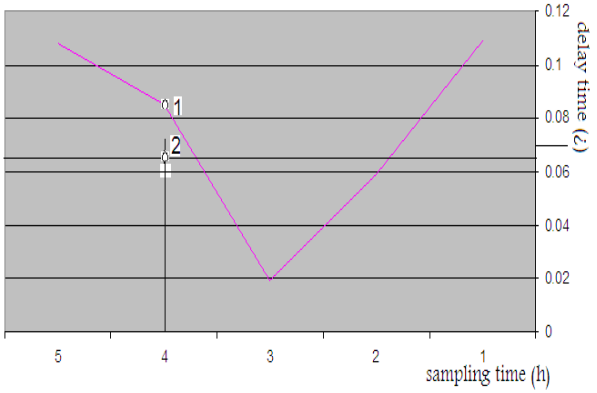


Fig. 4. A=–11, k=32 Primary Rough Relation Before Correction

In Figure 4, the relation should be piece wised to give the results shown in Table 5

TABLE V. SAMPLING TIME AGAINST THE DELAY TIME

h, sec.	y	, sec.τ
0.001	12	0.109
0.002	10	0.059
0.003	8.5	0.019
0.004	7	0.063
0.005	6	0.108

The least square root method is used as in the first case, so that the constant value is calculated.

a=0.21, b=–117.23, c=19571

i.e., at A=–11, k=32, the quadratic equation is

$$0.21 - 117.23\,h + 19571\,h^2 = 0 \quad (13)$$

Differentiate the previous equation

$$\frac{d\tau}{dh} = -117.23 + 3914h = 0$$

Thus, h=0.003, τ =0.034

The lowest point is (0.003, 0.034).

These values are substituted in equation 13, and the corrected delay time is shown in Table 6.

TABLE VI. SAMPLING TIME AGAINSTTHE DELAY TIME

h sec	secτ
0.001	0.113
0.002	0.054
0.003	0.034
0.004	0.054
0.005	0.113

Figure 5 depicts the relation between the corrected sampling time and the delay time after the optimized point according to h=3 m sec.

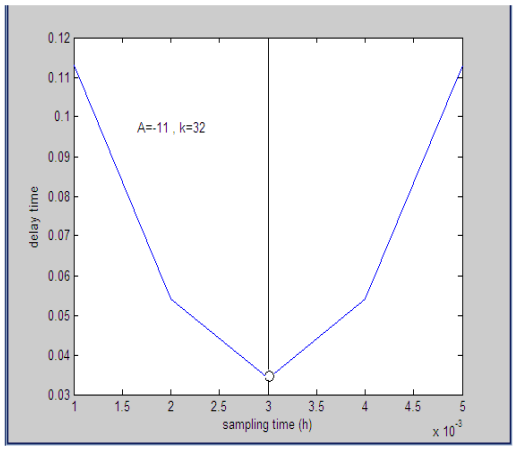


Fig. 5. Final Sampling Time Againstthe Delay Time (A=–11, k=32)

Comparing Figure 4 and Figure 5 shows that the system performance after correction is improved, and an optimization point corresponds to the sampling time, at h=0.003 sec.

V. CONCLUSION

In this paper, the following points can be concluded: (1) The sampling time and the delay time are optimized to improve the performance of a control system. (2) Two different operating points are considered to assess the efficiency of the proposed approach in improving the control system performance. (3) The point of the minimum delay time should be studied to assess the system performance. (4) When the sampling time becomes close to zero, the performance will be close to that of the continuous system. (5) The equation proves that when the sampling time increased the delay time increased, and vice versa.

REFERENCES

[1] Yepez, J.; Marti, P.; Fuertes, J.M., "Control loop performance analysis over networked control systems," IECON 02 [Industrial Electronics Society, IEEE 2002 28th Annual Conference of the] , vol.4, no., pp.2880,2885 vol.4, 5-8 Nov. 2002

[2] Lynch, C.B.; Dumont, G.A., "Control loop performance monitoring," Control Systems Technology, IEEE Transactions on , vol.4, no.2, pp.185,192, Mar 1996

[3] Grimble, M. (2002). Restricted Structure Control Loop Performance Assessment for State-Space Systems, in Proceedings of the 2002 American Control Conference vol.2, no., pp.1633,1638 vol.2.

[4] Zhang Tong; Wang Qinglin, "Linear Time-Variant Multivariable Feedback Control Loop Performance Assessment," Control Conference, 2007. CCC 2007. Chinese , vol., no., pp.215,219, July 26 2007-June 31 2007

[5] Yook, J.K.; Tilbury, D.M.; Soparkar, N.R., "A design methodology for distributed control systems to optimize performance in the presence of time delays," American Control Conference, 2000. Proceedings of the 2000 , vol.3, no., pp.1959,1964 vol.3, 2000

[6] Joelianto, E.; Sutarto, H.Y.; Wicaksono, A., "Compensation of delay time uncertainties on industrial control ethernet networks using LMI based robust H∞ PID controller," Wireless and Optical Communications Networks, 2008. WOCN '08. 5th IFIP International Conference on , vol., no., pp.1,5, 5-7 May 2008

[7] Samaranayake, L.; Leksell, M.; Alahakoon, S., "Relating Sampling Period and Control Delay in Distributed Control Systems," Computer as a Tool, 2005. EUROCON 2005.The International Conference on , vol.1, no., pp.274,277, 21-24 Nov. 2005

[8] Park, J., Han, S., and Kim, J. (2014). Improvement of Shape Recognition Performance of Sendzimir Mill Control Systems Using

Echo State Neural Networks, International Journal of Iron and Steel Research, 21(3):321-327.

[9] Mikulandri, R., Loncar, D., Cvetinovic, D., and Spiridon, G. (2013). Improvement of Existing Coal Fired Thermal Power Plants Performance by Control Systems Modifications, Journal of Energy, 57:55-65.

[10] Maki, M.; Hagino, K., "Adaptive performance improvement in a robust control system," Decision and Control, 2001. Proceedings of the 40th IEEE Conference on , vol.2, no., pp.1547,1548 vol.2, 2001

[11] Antunes, A.; Dias, F.M.; Vieira, J. & Mota, A. (2008) A neuro-fuzzy delay compensator for distributed control systems. Proceedings of the IEEE Int. Conf. on Emerging Technologies and Factory Automation, Hamburg, Germany, September, 2008, pp. 1088-1091.

[12] Robert Mikulandri, Dra_zen Lon_car, Dejan Cvetinovi, Gabriel Spiridon**, "** Improvement of existing coal fired thermal power plants performance by control systems modifications**",** Energy 57 (2013) 55e65**.**

[13] Xi-Sheng Zhan, Zhi-HongGuan, Fu-ShunYuan, Xian-HeZhang, "Optimal performance of discrete-time control systems based on network-induced delay", European Journal of Control 19 (2013) 37–41.

[14] Changchun Hua , LiuliuZhang, Xinping Guan, " Output feedback control for interconnected time-delay systems with prescribed performance", Neurocomputing129(2014)208–215.

[15] Song Han ; Kam-Yiu Lam ; Jiantao Wang ; Ramamritham, K. ;Mok, A.K. , "On Co-Scheduling of Update and Control Transactions in Real-Time Sensing and Control Systems: Algorithms, Analysis, and Performance", IEEE Transactions on Knowledge and Data Engineering, Vol. 25 ,, (2013), pp. 2325-2342.

[16] Meng, D. ; Jia, Y. ; Du, J. ; Yu, F. , " Robust iterative learning control design for uncertain time-delay systems based on a performance index", Control Theory & Applications, IET Volume: 4 , (2010), pp. 759-772.

[17] Cea, M. ; Salgado, M.E. "Performance bounds on the control of highly sparse discrete-time multivariable systems", Control Theory & Applications, IET, Vol. 4 , (2010), pp. 1319-1329.

[18] Shen Yin ; Hao Luo ; Ding, S.X. , " Real-Time Implementation of Fault-Tolerant ControlSystems With Performance Optimization", IEEE Transactions on Industrial Electronics, Vol. 61 , Issue: 5, (2014), pp. 2402-2411.

[19] Katsuhiko Ogata. 2001. Modern Control Engineering (4th ed.). Prentice Hall PTR, Upper Saddle River, NJ, USA.

[20] Etschberger, K. (2001). Controller Area Network, Basics, Protocols, Chips and Applications, IXXAT Automat, Gmbh, German.

[21] Kopetz, H., "Sparse time versus dense time in distributed real-time systems," Distributed Computing Systems, 1992., Proceedings of the 12th International Conference on , vol., no., pp.460,467, 9-12 Jun 1992

Use of Non-Topological Node Attribute Values for Probabilistic Determination of Link Formation

Abhiram Gandhe
Computer Science and Engg.
Visvesvaraya National Institute of Technology
Nagpur, India

Parag Deshpande
Computer Science and Engg.
Visvesvaraya National Institute of Technology
Nagpur, India

***Abstract*—Here we propose a probabilistic model for determining link formation, using Naïve Bayes Classifier on non-topological attribute values of nodes, in a social network. The proposed model gives a score which helps to determine the relationship strength in a non-formed link. In addition to Naïve Bayes Classifier, weighted Average of the Attribute value match helps to determine the friendship score of a non-formed link.**

With the increase in online social networks and its influence on people, more and more individuals are getting wider and enhanced social connect. Everyone tries to connect more to explore more. In this race of more, an individual needs better and definitive tools to help them grow their network. Wider is the network more is the possibility to explore.

Here we present a novel approach for predicting a link (friendship) between two individuals (nodes) in a social network. The proposed approach uses non-topological attribute data values of both the nodes and predicts linkage possibility by applying Naïve Bayes Classifier on non-topological attribute data values of nodes in existing linkages.

A linkage possibility is expressed using one quantitative measure FSCORE. We call it friendship score (FSCORE) between two unconnected individuals. FSCORE is used to predict linkage between two nodes. Higher FSCORE means a higher possibility of linkage between two nodes.

Keywords*—*Non-Topological Attribute; Link Prediction; Naïve Bayes Classifier; Weighted Average; Graph Database; Social Network; Data Mining

I. INTRODUCTION

Online Social Networks (OSNs) have become an integral part of today's life. OSN is where everyone keeps her/his social connect. Social networks are still doing the same work like information exchange, furthering a cause, keeping up communication, guiding and developing a society, to name a few. More connected is an individual; more he can achieve out of his social connects.

In the early days greeting and meeting people in social gathering was the only way to increase your social network and influence. Today, with the acceptance and spread of online social networks, ways to connect individuals have significantly improved. Different Online Social Networks addresses different interests of an individual. Facebook and Google+ mainly exist to share information, initiate a conversation and discuss on a certain topic. Twitter a micro-blogging site helps commenting on any issue at hand/in mind and letting the world know about it. Flickr is a photo sharing social network. LinkedIn is an online network of professionals.

With every online Social Network site, there is a new and different social network, of people in a context, created by an individual. With the increasing options and different focus areas of different networks, the data created out of these networks is diverse and huge. This data provides a great opportunity for analysts to dig the created data and interpret the future course of the network.

While there are inherent risks in use and distribution of OSNs data, there are also many potential benefits of this data. Interpretation of Social networking data along with related tools created to interpret the data can help to strengthen existing relationships and provide opportunities for creating new relationships. With better means and tools, a stronger and more connected network can be intended and created. Today networks are deploying different techniques to help a user to grow their social circle, connect with new individuals and find superior content of interest.

Every individual is on the lookout to increase his network with people of interest. As faced by every individual, there are two impediments in connecting with individuals of interest.

1) Who is the one of interest?

2) How high is the possibility of connecting with the one of interest?

Different Social Networks are engaging different ways to augment a user's arsenal to help them grow their own network. Most common ways of predicting a higher probability of connecting in a network are:

a) Individuals with maximum mutual friends are suggested a connect

b) Individuals are asked to suggest a connect between their unconnected friends

c) Unconnected individuals having multiple short length paths in the graph are suggested a connect

d) Unconnected individuals commenting on the same conversation, multiple times are suggested a connect

New and better tools are evolving at day end to provide users with better services to enhance their experience of social connect. There is a wide range of research going on in the area of suggesting connects. In research terminology, it is called as Link Prediction in Graphs. Link prediction can be used to

identify hidden links, not yet formed in an Online Social Network, in a friend suggestion mechanism.

Link prediction outside the social network domain can have multiple uses like:

a) Recommendation and relevance prediction in e-commerce [3]

b) Protein Interaction prediction in Life Sciences [2]

c) Identifying hidden groups of terrorist or criminals using link prediction in the security domain [4]

The link prediction problem is relevant to different scenarios; several algorithms have been proposed in recent years to solve it. One common approach for solving **Link prediction** problem is using **supervised learning** algorithm. This approach was introduced by Liben-Nowell and Kleinberg in 2003 [6], who studied the usefulness of graph topological features by testing them on bibliographic data sets. In 2006, the work was extended to identifying hidden group of terrorist by Hasan et al [4] and since then several other researchers have implemented this approach. Most of the solutions, that these researchers proposed were tested on bibliographic or onco-authorship data sets [4], [6], [7], and [8]. In 2009, Chen et al [1] depicted several algorithms used by IBM on their internal social network, which enable its employees to connect with each other. Song et al. used matrix factorization to estimate the similarity between nodes in real life social networks such as Facebook and MySpace [9]. In 2011, W J. Cukierski et al [10] extracted 94 distinct graph features. Using a Random Forests classifier, they achieved impressive results in predicting links on Flickr datasets.

Here we are proposing a novel approach for predicting a link (friendship) between two individuals (nodes) in a social network using OSN (Online Social Network) data and predict linkage possibility by applying Naïve Bayes Classifier on attribute data values of nodes in existing linkages.

II. Problem Statement

Classification of links in social network can be done on different types of node data:

a) Topological Attribute Data

b) Node Interaction Data

c) Non-Topological Attribute Data

All the above types of node data can be used to classify the links. The classification helps in predicting the possibility of connection between two non-connected nodes. Most of the research to date is done on Topological Attribute data and Node interaction Data.

In this paper, we propose the **mechanism of classification based on non-**topological **attribute data**. The dataset used for experimentation will be from Facebook™. We will be using Naïve Bayes classifier for classifying the existing links and use the classification for predicting a link between two nodes.

III. Methodology

A. Online Social network

The Social Network in consideration, in this paper, is Facebook. Facebook is an online social interaction and networking service. A user above 13 years of age can create an account on Facebook. On Facebook, a user can make friends with other Facebook users. A user can post anything on her/his wall (representation of profile space) or her/his friend's wall. A user can "Like" or "comment" on posts by her/him or her/his friends. The posts, "Like" action and comments can be termed as public interaction between users. All the public interactions between users, done by a user, are available for view to all users on the timeline of the user.

Other than public interaction a user can have private interaction with a friend user. The possible ways of private communication is chatting or inbox messaging. All the private communications are confidential and are visible only to the two users between whom the interaction has taken place. These public and private interactions between users are termed as node interaction data and can be used to predict friendship.

A user stores his profile information as the time of registration with Facebook. Profile information on Facebook can range from First Name, Last Name, Date of Birth, Gender, Religion, Home Town, Current City and Relationship Status to Work, Work History and Education History information. Over the period of time, a user can update, add or delete profile information. A user can also put restrictions on visibility of this information from "Public" to "Friends Only" to "Only Me" to any other specific friends group available. Due to the selective visibility of data governed by the user and nearly all the attributes are optional there is a wide possibility of having attribute values as blank.

B. DatasetPreparation

A sample subset of Facebook was used as a dataset to work on. This Dataset was extracted from Facebook using Facebook App named "FBNetworkAnalysis". URL for this app is "https://app.facebook.com/mytestappfbabhi". The data was collected from 7637 users. In the context of this analysis, users are represented as nodes and friends are represented as two nodes on an edge. A friendship is represented as a link between two users. A user can mark a link as "Friend", "Cousin" (any other relative) or "Spouse"/"Significant Other". This Relationship is taken as name/type of the link. Link name/type is not considered in this analysis.

All the profile information made available by the user are considered as the node attributes and the analysis of the links is done using the values of the node attributes of the users (nodes) in a link (edge).

The values extracted using the Facebook application:

Date of Birth	**Gender**	Religion
Home Town	**Current City**	Relationship Status
Interested In	College/**School**	Education Year
Work Company	Work Location	Work Year
Favorite Athlete	Favorite Team	

C. Data Representation

fbgraph (**finite**, **no multiedges**, **undirected**)

Facebook sample data set will be represented as a graph with finite nodes and a finite number of connections. A

connection (Link) can only be established when friend request is sent by one user (Node) and accepted by another user (Node).Mutual acceptance by both the **nodes** makes the **link undirected**. There is only one link between two nodes, the link type can differ on the nodes in consideration ("Friend", "Relative" or "Spouse") so there will be no multi-edges.

fbG = (U, L) where

U (or U (fbG)) is a set of nodes

L (or L (fbG)) is a set of links, each of which is a set of two nodes (undirected)

Two nodes that are associated with a link are adjacent nodes.

Let n = |U| and m = |L|

The neighbor of each node u is

N (u) = {v |uvƐ L}

The degree of user u is d (u) = |N (u)|

D. Naïve Bayes Classifier

Naïve Bayes classifier depends on Bayes theorem

$$\boldsymbol{p(cj|d)} = \frac{\boldsymbol{p\left(\frac{d}{cj}\right)p(cj)}}{\boldsymbol{p(d)}}$$

Where,

$\boldsymbol{p(cj|d)}$: Probability of instance **d** being in class **cj**
(This is what we will be computing)

$\boldsymbol{p(d/cj)}$: Probability of generating instance **d** given class **cj**
*(We can imagine that being in class **cj**, causes you to have feature **d** with some probability)*

$\boldsymbol{p(cj)}$: Probability of occurrence of class **cj**
*(This is just how frequent the class **cj**, is in our database)*

$\boldsymbol{p(d)}$: Probability of instance **d** occurring
*(This is just how frequent the instance **d**, is in our database)*

And

$$\boldsymbol{p(cj/d)} = \frac{\boldsymbol{p(cj \cap d)}}{\boldsymbol{p(d)}}$$

Where,

$\boldsymbol{p(cj|d)}$: Probability of instance **d** being in class **cj**

$\boldsymbol{p(cj \cap d)}$: Existing Links having instance **d** in class **cj**

$\boldsymbol{p(d)}$: Probability of instance **d** occurring

To simplify the task, **naïve Bayesian classifiers** assume attributes have independent distributions, and thereby estimate:

$$\boldsymbol{p(d|cj)} = \boldsymbol{p(d1|cj)} \times \boldsymbol{p(d2|cj)} \times \ldots \times \boldsymbol{p(dn|cj)}$$

d: instance d

d1: Value of feature 1

d2: value of Feature 2

so on and so forth …

Advantages of Naïve Bayes Classifier:

- ✓ Fast to train (single scan). Fast to classify
- ✓ Not sensitive to irrelevant features
- ✓ Handles real and discrete data
- ✓ Handles streaming data well

Disadvantages of Naïve Bayes Classifier:

- ➢ Assumes independence of features

E. Data Analysis

Function	Equation	Number
Number of Users		7637
Possible Friend Connections		29158066
Existing Friend Connections	P(Friend)	101904
Non Friend Pairs	P(Non Friend)	29056162
Possible Friends with Same Gender	P(Same Gender)	12682300
Friends of Same Gender	P(Same Gender ∩ Friend)	60604
Non Friends of Same Gender	P(Same Gender ∩ Non Friend)	12621696
Possible Friends with Different Gender	P(Different Gender)	9094400
Friends of Different Gender	P(Different Gender ∩ Friend)	35379
Non Friends of Different Gender	P(Different Gender ∩ Non Friend)	9059021
Possible Friends with Same Location	P(Same Location)	2524621
Friends of Same Location	P(Same Location ∩ Friend)	27013
Non Friends of Same Location	P(Same Location ∩ Non Friend)	2497608
Possible Friends with Different Location	P(Different Location)	9917945
Friends of Different Location	P(Different Location ∩ Friend)	36839
Non Friends of Different Location	P(Different Location ∩ Non Friend)	9881106
Possible Friends with Same School	P(Same School)	726648
Friends with Same School	P(Same School ∩ Friend)	17117
Non Friends with Same School	P(Same School ∩ Non Friend)	709531
Possible Friends with Different School	P(Different School)	11388855
Friends with Different School	P(Different School ∩ Friends)	36977
Non Friends with Different School	P(Different School ∩ Non Friend)	11351878
Possible Friends with Same Favorite Athlete	P(Same Favorite Athlete)	866676
Friends with Same Favorite Athlete	P(Same Favorite Athlete ∩ Friend)	4803
Non Friends with Same Favorite Athlete	P(Same Favorite Athlete ∩ Non Friend)	861873
Possible Friends with Different Favorite Athlete	P(Different Favorite Athlete)	1211065
Friends with Different Favorite Athlete	P(Different Favorite Athlete ∩ Friend)	7836
Non Friends with Different Favorite Athlete	P(Different Favorite Athlete ∩ Non Friend)	1203229

F. Conditional Probability Bayes Rule

Using Conditional Probability Bayes Rule on the above data:

$$p(cj/d) = \frac{p(cj \cap d)}{p(d)}$$

Same Gender:

$$P\left(\frac{Friend}{Same\ Gender}\right) = \frac{P(Friend \cap Same\ Gender)}{P(Same\ Gender)}$$

$$= \frac{60604}{12682300} = 0.0047786$$

Different Gender:

$$P\left(\frac{Friend}{Different\ Gender}\right) = \frac{P(Friend \cap Different\ Gender)}{P(Different\ Gender)}$$

$$= \frac{35379}{9094400} = 0.0038902$$

Same Location:

$$P\left(\frac{Friend}{Same\ Location}\right) = \frac{P(Friend \cap Same\ Location)}{P(Same\ Location)}$$

$$= \frac{27013}{2524621} = 0.0106998$$

Different location:

$$P\left(\frac{Friend}{Different\ Location}\right) = \frac{P(Friend \cap Different\ Location)}{P(Different\ Location)}$$

$$= \frac{36839}{9917945} = 0.0037144$$

Same School:

$$P\left(\frac{Friend}{Same\ School}\right) = \frac{P(Friend \cap Same\ School)}{P(Same\ School)}$$

$$= \frac{17117}{726648} = 0.0235561$$

Different School:

$$P\left(\frac{Friend}{Different\ School}\right) = \frac{P(Friend \cap Different\ School)}{P(Different\ School)}$$

$$= \frac{36977}{11388855} = 0.0032468$$

Same Favorite Athlete:

$$P\left(\frac{Friend}{Same\ Favorite\ Athlete}\right) = \frac{P(Friend \cap Same\ Favorite\ Athlete)}{P(Same\ Favorite\ Athlete)}$$

$$= \frac{4803}{866676} = 0.0055419$$

Different Favorite Athlete:

$$P\left(\frac{Friend}{Different\ Favorite\ Athlete}\right) = \frac{P(Friend \cap Different\ Favorite\ Athlete)}{P(Different\ Favorite\ Athlete)}$$

$$= \frac{7836}{1211065} = 0.0064703$$

G. Attribute Value Weightage

Calculated conditional probability above shows that two individuals in a same school have a higher probability of being friends than being from the same location.

Putting this probability as weight for the attribute values of two nodes, following are the weights with the maximum weight attribute at the top.

Attribute Value Weights:

W (Same School)	**0.0235561**
W (Same Location)	**0.0106998**
W (Different Favorite Athlete)	**0.0064703**
W (Same Favorite Athlete)	**0.0055419**
W (Same Gender)	**0.0047786**
W (Different Gender)	**0.0038902**
W (Different Location)	**0.0037144**
W (Different School)	**0.0032468**

If two individuals have same school populated in the School attribute the weight of having a friendship will be **0.0235561** if school populated in the School attribute for both the individuals is different, then weight of having a friendship will be **0.0032468** instead.

If any of the individuals does not have the school attribute populated/shared, then no weight is added for school attribute in the friendship score.

H. Friendship Score (FSCORE)

If a Link is present between two nodes, a and b, then:

$$\mathbf{FSCORE}[\mathbf{a}, \mathbf{b}] = \mathbf{1} + \mathbf{FNT}(\mathbf{a}, \mathbf{b})$$

Where FNT (a,b) is a function non-topological attribute of a and b. If a link is not formed between two nodes, a and b, then friendship score needs to be calculated using non-topological attribute data. The FSCORE is calculated as:

$$\mathbf{FSCORE}[\mathbf{a}, \mathbf{b}] = \mathbf{FNT}(\mathbf{a}, \mathbf{b})$$

If two individuals (Non Friends) have **only Same Gender** and no other attributes populated, the probability of Friendship is 0.0047786 and hence the Friendship Score for future friendship is **0.0047786.** Similarly, if two individuals **only** have **Same Location** then their Friendship Score will be **0.0106998**.

IV. Mathematical Model for Dataset with Missing Data

FSCORE or Attribute weight calculated here is with the data where there are missing attribute values in many node elements. Also in the case of the social data in consideration, all features/attributes cannot be assumed to be independent of each other. Considering features are dependent on each other

Naïve Bayesian distributed probability equation cannot be used here. Instead, we propose the use of the weighted average.

$$\mathbf{FNT}(\mathbf{a},\mathbf{b}) = \frac{\sum_{i=1}^{n}\sum_{j=1}^{m} \boldsymbol{Wij} \times \boldsymbol{Vij}}{\boldsymbol{n}}$$

Where:

i: Attributes/Features e.g. Gender (if there is no value associated with the attributes in any/both of the nodes **a** or/and **b** then that attribute will not be taken up for the calculation of FSCORE)

j: Different Attribute Values possible e.g. Same Gender, Different Gender (No gender available is also a valid possible value, but it will already be excluded from the equation because of the elimination of attributes while collating the final set of i's)

According to the above, FSCORE for two nodes with Same Gender (SG), Same Location (SL), Same School (SS) and Same Favorite Athlete (SFA), and no other attribute value populated, in a network where there are Lots of nodes with missing attribute values is as follows. For two unconnected nodes:

$$\text{fscore}(\text{SG},\text{SL},\text{SS},\text{SFA}) = \text{FNT}(\text{SG},\text{SL},\text{SS},\text{SFA})$$

$$= \frac{\begin{array}{c} WSG \times VSG + WDG \times VDG + \\ WSL \times VSL + WDL \times VDL + \\ WSS \times VSS + WDS \times VDS + \\ WSFA \times VSFA + WDFA \times VDFA \end{array}}{4}$$

$$= \frac{\begin{array}{c} WSG \times 1 + WDG \times 0 + WSL \times 1 + WDL \times 0 \\ +WSS \times 1 + WDS \times 0 + WSFA \times 1 + WDFA \times 0 \end{array}}{4}$$

$$= \frac{WSG + WSL + WSS + WSFA}{4}$$

$$= \frac{0.0235561 + 0.0106998 + 0.0055419 + 0.0047786}{4}$$

$$= \frac{0.0445764}{4} = \mathbf{0.0111441} \quad (1)$$

Here,

WSG: Weightage for Same Gender (weight to be added to FSCORE if the two nodes under consideration have the same gender)

Similarly WDG, WSL, WDL, WSS, WDS, $WSFA$, $WDFA$ are weights for Different Gender, Same Location, Different Location, Same School, Different School, Same Favorite Athlete and Different Favorite Athlete, respectively.

VSG: Value in gender attribute for both the nodes is same. VSG = 1 if both the nodes under consideration have a same gender and VSG = 0 if gender is different for both nodes or any of the node doesn't have gender value

Similarly VDG, VSL, VDL, VSS, VDS, $VSFA$, $VDFA$ are values for Different Gender, Same Location, Different Location, Same School, Different School, Same Favorite Athlete and Different Favorite Athlete, respectively.

V. Mathematical Model for Dataset with No Missing Data

When no nodes are missing attribute data, every attribute value matches or does not match between two nodes. In such cases calculating FSCORE can be done differently

In the Training set, number of friends with:

Same Gender + Same Location: **18154**
Same Gender + Different Location: **21524**
Different Gender + Same Location: **8379**
Different Gender + Different Location: **15240**

Friendship Score for Same Gender and Same Location:

$$fscore\ (SG, SL) = \frac{18154}{(18154 + 21524 + 8379 + 15240)}$$

= **0.2868066**

FSCORE for two unconnected nodes with Same Gender (SG), Same Location (SL), Same School (SS) and Same Favorite Athlete (SFA), in a network where there are no nodes with missing attribute values is as follows.

In the Training set, Number of Friends with:

SG + SL + SS + SFA = **381**
SG + SL + SS + DFA = **572**
SG + SL + DS + SFA = **612**
SG + SL + DS + DFA = **964**
SG + DL + SS + SFA = **423**
SG + DL + SS + DFA = **237**
SG + DL + DS + SFA = **447**
SG + DL + DS + DFA = **524**
DG + SL + SS + SFA = **144**
DG + SL + SS + DFA = **208**
DG + SL + DS + SFA = **212**
DG + SL + DS + DFA = **1102**
DG + DL + SS + SFA = **68**
DG + DL + SS + DFA = **83**
DG + DL + DS + SFA = **208**
DG + DL + DS + DFA = **310**

$$\text{fscore}(\text{SG},\text{SL},\text{SS},\text{SFA}) = \text{FNT}(\text{SG},\text{SL},\text{SS},\text{SFA})$$

$$= \frac{381}{\begin{array}{c}(381 + 572 + 612 + 964 + 423 \\ +237 + 447 + 524 + 144 + 208 \\ +212 + 1102 + 68 + 83 + 208 + 310)\end{array}}$$

$$= \mathbf{0.0586605} \quad (2)$$

The $\text{fscore}(\text{SG},\text{SL},\text{SS},\text{SFA})$ in 1 is different than the one in 2 due to the difference in dataset. Dataset used in equation 2 is the subset (nodes with no missing values for gender, location, school and favorite athlete attributes) of the one used for equation 1.

In the case, if complete data is available, number of permutation combination to store and update, on link formation, increases with the increase in the attributes in consideration. This becomes cumbersome to maintain and update the data of all the combinations of attributes. In the case

of n attributes, 2^n combinations need to be maintained. For excluding some of the attributes from the final set of test, attributes may lead to maintaining different combinations separately.

What we propose for this is an equation of approximation.

$$p(A \cap B \cap C) = p(A|B)\, p(B|C)\, p(C)$$

On the same data set used in equation 2:

$$p(SG \cap SL \cap SS \cap SFA)$$

$$= p(SG|SL)\, p(SL|SS)\, p(SS|SFA)\, p(SFA)$$

$$= \frac{2529}{4195} \times \frac{1305}{2116} \times \frac{1016}{2495} \times \frac{2495}{6495}$$

$$= 0.6028605 \times 0.6167297 \times 0.4072144 \times 0.3841416$$

$$= \mathbf{0.0581603} \qquad \text{-(3)}$$

We consider a relation of two attributes and use the relations in a link to the next attribute in consideration. What we have done in above example is have a probability of Same Gender for Same Location with Probability of Same Location for Same School and Probability of Same School for Same Favorite Athlete along with Probability of a link having Same Favorite Athlete.

This is done on the data set which has complete data and no Missing Values. The nearness of the FSCORE in equation 2 and 3 confirms approximation works well with proposed formula.

VI. Conclusions and Recommendation

A relationship is made on different parameters and we have tried to quantify the parameters for relationship building, depending on an existing link/relationship data as stated in the paper. Deriving a possibility of a relationship (FSCORE) can be analyzed using the proposed model in this paper.

FSCORE is an effective way of predicting the possibility of relationship/link between two nodes using Non-topological attribute values of nodes. Significance and weight, of non-topological attributes, is determined by the already existing links and recurrence of a value pattern for these non-topological attributes in existing links.

FSCORE can be used to calculate the cost of connecting to a distant node in a graph. FSCORE can provide a measure of strength between two unconnected nodes in order to make decisions or predictions in a different set of problems in a graph network. FSCORE can also provide a factor to help identify/quantify connected nodes. FSCORE can be used to compare and rate a relation of connected or unconnected nodes stronger or weaker to other relations.

In a graph network, if a link of reference is to be invoked or an optimized path for traversal has to be identified, then FSCORE can provide a quantitative value for analysis between two connected or unconnected node. FSCORE can be used as a relationship cost parameter in similar Graph Network problems. FSCORE is calculated using non-topological attribute values between nodes and can be coupled with topological attribute data to improve the prediction possibility.

References

[1] J. Chen, W. Geyer, C. Dugan, M. Muller, and I. Guy, "Make new friends, but keep the old: recommending people on social networking sites," in Proceedings of the 27th international conference on Human factors in computing systems, ser. CHI '09. NewYork, NY, USA: ACM, 2009, pp. 201–210. (*references*)

[Online]. Available:http://doi.acm.org/10.1145/1518701.1518735

[2] E. M. Airoldi, D. M. Blei, S. E. Fienberg, and E. P. Xing, "Mixed membership stochastic block models for relational data with application to protein-protein interactions," Proceedings of International BiometricSociety-ENAR Annual Meetings, 2006.

[3] Z. Huang, X. Li, and H. Chen, "Link prediction approach to collaborative filtering," Proceedings of the 5th ACM/IEEE-CS joint conference onDigital libraries, 2005.

[4] M. A. Hasan, V. Chaoji, S. Salem, and M. Zaki, "Link prediction using supervised learning," SDM Workshop on Link Analysis, Counterterrorism and Security, 2006.

[5] M. A. Hasan and M. J. Zaki, Social Network Data Analytics, C. C.Aggarwal, Ed. Springer, 2011.

[6] D. Liben-Nowell and J. Kleinber, "The link-prediction problem for social networks," Journal of the American Society for Information Science andTechnology, vol. 58, no. 7, 2007.

[7] J. R. Doppa, J. Yu, P. Tadepalli, and L. Getoor, "Chance-constrained programs for link prediction," In Proceedings of Workshop on AnalyzingNetworks and Learning with Graphs at NIPS Conference, 2009.

[8] H. R. Sa and R. B. C. Prudencio, "Supervised learning for link prediction in weighted networks," III International Workshop on Web and TextIntelligence, 2010.

[9] H. H. Song, T. W. Cho, V. Dave, Y. Zhang, and L. Qiu,"Scalable proximity estimation and link prediction in online social networks," in Proceedings of the 9th ACM SIGCOMMconference on Internet measurement conference, ser. IMC '09. NewYork, NY, USA: ACM, 2009, pp. 322–335.

[10] [Online]. Available:http://doi.acm.org/10.1145/1644893.1644932

[11] W. J. Cukierski, B. Hamner, and B. Yang, "Graph-based features for supervised link prediction," International Joint Conference on NeuralNetworks, 2011.

Different Classification Algorithms Based on Arabic Text Classification: Feature Selection Comparative Study

Ghazi Raho
MIS Dept./ Amman Arab University
Amman-Jordan

Riyad Al-Shalabi
MIS Dept./ Amman Arab University
Amman-Jordan

Ghassan Kanaan
CS Dept./ Amman Arab University
Amman-Jordan

Asma'aNassar
CS Dept./ JUST University
Irbid-Jordan

***Abstract*—Feature selection is necessary for effective text classification. Dataset preprocessing is essential to make upright result and effective performance. This paper investigates the effectiveness of using feature selection. In this paper we have been compared the performance between different classifiers in different situations using feature selection with stemming, and without stemming.Evaluation used a BBC Arabic dataset, different classification algorithms such as decision tree (D.T), *K*-nearest neighbors (*K*NN), Naïve Bayesian (NB) method and Naïve Bayes Multinomial(NBM) classifier were used. The experimental results are presented in term of precision, recall, F-Measures, accuracy and time to build model.**

Keywords—Text Classification; Feature Selection; Arabic Text; Recall; F-Measure

I. INTRODUCTION

We know that the amount of Arabic information that founded on the internet is very large and increasing rapidly. This growth directs researchers to find some of the effectiveness mechanism and good tools that may help the researchers to better managing, filtering, processing and classification a large Arabic information resource. Text classification (TC) is the task using to classify a specific dataset into different classes; it also called document classification, text categorization or document categorization.

TC also used to solve some research problems such as information retrieval (IR), data mining, and natural language processing. There are many applications on TC like document indexing, document organization, text filtering, word sense disambiguation, speech recognition and web text hierarchical categorization.

TC can use as a binary classification like -nearest neighbors (KNN), Naïve Bayesian method and SVM and as a multi classification like boosting and multi-class SVM.

TC task can divides the dataset into two part: training set and testing set, the classifier algorithm learn on training to build a TC model, then TC system to classify the testing set into different classes, To achieve effective performance we used feature selection methods.

To get a better performance wedid some preprocessing steps on the dataset which we will talk about later in this paper. Section two will talk about the related work, section three will talk about our objectives, section four talk about experimental results, and then conclusion and future work, and finally the references.

II. RELATED WORK

In [1] the authors presented the performance of using a Support Vector Machines (SVMs) based text classification system on Arabic text. The authors using one of the feature selection methods which is CHI square method, theyuseda preprocessing stepsin their work to give a better evaluation. The proposed system gives good results. To classify any text we must determine a set of features to achieve best classification. This paper presents the effectiveness of six features selection method to extract and choose a good features from Arabic document. The authors used SVM classifier algorithm to compare the performance between these six methods (CHI, NGL, GSS, IG, OR and MI).

The authors in [2] used an in-house collected corpus from online Arabic newspaper archives, including Al-Jazeera, Al-Nahar, Al-Hayat, Al-Ahram, and Al-Dostor. The collected corpus consists of 1445 documents. These documents consist of nine categories, the authors did some Pre-processing for the dataset such as remove digits and punctuation marks, all the non-Arabic texts were filtered, remove the Arabic function words (stop words) and other. In [2] the result showing that CHI, NGL and GSS performed most effective with SVMs for Arabic TC tasks, but OR and MI performed terribly. In [3] the authors talked about three contributions: (i) showing successful classification of Arabic documents, (ii) make their database available to other researchers, (iii) find a better performancebetween Binary PSO and K-nearest neighbor using feature selection methods. In [3] the authors presented BPSO - KNN as a feature selection method and applied this method on three Arabic text dataset. The authors used three classification algorithms which are SVM, Naïve Bayes andC4.5 decision tree learning.

In [4] the authors used Chi-Squaremethod as a pre-processing step which applied on dataset before doing the classification. In [4] the authors compared between the proposed method and other feature selection methods the result shows that the proposed methodperformed better performance than other features selection methods.

III. Our Objectives

To compare the performance between different classification algorithm (decision tree, K-nearest neighbors(KNN), Naïve Bayesian method and Naïve Bayes multinomial classifier) in different situations: using feature selection methods with light stemmer, (khoja stemmer) and using feature selection with full word.

A. TC Process

Text classification system usually separated into three main phases which are : *Data preprocessing and feature selection phase that* makes the dataset more compatible and applicable to train the text classifier, *text classifier phase* that use to classify dataset into different classes, *and evaluation phase*to show the performance of the used classification algorithm.

B. Arabic Dataset Preprocessing

There are a lot of Arabic dataset available on the internet that can be used, we used BBC Arabic dataset that contains 4763 documents belongs to seven categories (News Middle East in 2356, News of the world in 1489, the economy and business 296, Sport 219, the press world 49, Science and Technology 232 Arts & Culture, 122). The dataset contains 1,860,786 words and 106,733key word. These dataset are processed according to the following steps:

1) Remove digits, dash, punctuation marks and any other mark.

2) Filtered all non-Arabic text.

3) Remove stop words from the text document (such as "بدا/", "أحد", "أخر" and other stop words).

4) Use feature selection methods with stemmer and with full word.

C. Feature Selection Methods

Feature selection (FS) is a task to choose a subset feature from the original feature set, FS is widely used in TC task. FS consist of following steps:

1) Feature generation: in this step we generate a subset of feature by using some search process.

2) Feature evaluation: in this step we used some evaluation matrices to measure the goodness of selected features.

3) Feature validation: in this step we used a validation procedure to measure if the selected features are valid or not.

In this paper we used two feature selection methods the Information Gain (IG), and theX^2statistics (CHI) as shown in table 1.

TABLE I. FS Methods

CHI	$\frac{N.[P(t_k,c_i).P(\bar{t}_k,\bar{c}_i)-P(t_k,\bar{c}_i)P(\bar{t}_k,c_i)]^2}{P(t_k)P(\bar{t}_k)P(c_i)P(\bar{c}_i)}$
IG	$\sum_{c\in\{c_i,\bar{c}_i\}}\sum_{t\in\{t_i,\bar{t}_i\}} P(t/c).log\frac{P(t/c)}{P(t).P(c)}$

D. Text Classifier

In this paper we used different classifiersthese classifiers are: decision tree, K-nearest neighbors (KNN), Naïve Bayesian method and Naïve Bayes multinomial,we have compared between the performance of these classifier in different terms of categorization effectiveness. wedivided the dataset into two parts, one for the training, and the other for testing.

E. TC Evaluation Measure

We have evaluated the performance for the classifiers (decision tree, K-nearest neighbors(KNN), Naïve Bayesian method and Naïve Bayes multinomial) in terms of precision, recall, accuracy, F-Measures and time to build model as shown in equations 1, 2, and 3.

$$Pi=TPi/(TPi+FPi) \quad (1)$$

$$Ri=TPi/(TPi+FNi) \quad (2)$$

$$Fi=2PiRi/(Ri+Pi) \quad (3)$$

IV. TC Experimental Results

We have used two feature selection methods (CHI and IG), four classifiers (decision tree, K-nearest neighbors(KNN), Naïve Bayesian method and Naïve Bayes multinomial classifier) were used, a Weka tools of version 3.7 were used, the results are shown in table II to table X.

TABLE II. Khoja Stemmer Experiments by Taking Chi-Square Feature Selection Type

Classifier type	Time to build model/ sec	Chi-Square feature selection results			
		accuracy	Average Precision	Average recall	F-Measures
D.T	33.67	99.6221 %	0.996	0.996	0.996
NB	4.01	90.9091 %	0.932	0.909	0.917
KNN	0.01	73.1262 %	0.807	0.731	0.716
NBM	0.16	92.7357 %	0.935	0.927	0.928

TABLE III. KHOJA STEMMER EXPERIMENTS BY TAKING IG RATIO SELECTION FEATURE

Classifier type	Time to build model	Info Gain			
		Accuracy	Precision	Recall	F-Measures
D.T.	36.27	99.6221	0.01	0.99	0.996
NB	4.61	90.9091	0.93	0.91	0.917
KNN	0.01	73.1262	0.81	0.73	0.716
NBM	0.06	92.7357	0.94	0.93	0.928

TABLE IV. KHOJA STEMMER EXPERIMENTS BY TAKING NO FEATURE SELECTION TYPE

Classifier type	Time to build model	Null Feature Selection Type			
		Accuracy	Precision	Recall	F-Measures
D.T.	31.5	99.4751 %	0.995	0.995	0.995
NB	4.26	90.9091 %	0.932	0.909	0.917
KNN	0.01	73.1262 %	0.807	0.731	0.716
NBM	0.17	92.7357 %	0.935	0.927	0.928

TABLE V. LIGHT STEMMER EXPERIMENTS BY TAKING CHI SQUARE FEATURE SELECTION TYPE

Classifier type	Time to build model	Chi-square Feature Selection Type			
		Accuracy	Precision	Recall	F-Measures
D.T.	49.4	99.4961 %	0.995	0.995	0.995
NB	5.43	91.9169 %	0.931	0.919	0.922
KNN	0.01	66.3657 %	0.891	0.664	0.675
NBM	0.17	92.0638 %	0.927	0.921	0.921

TABLE VI. LIGHT STEMMER EXPERIMENTS BY TAKING IG RATIO FEATURE SELECTION TYPE

Classifier type	Time to build model	Info-gain ratio Feature Selection Type			
		Accuracy	Precision	Recall	F-Measures
D.T.	54.1	99.5591 %	0.996	0.996	0.996
NB	7.07	91.9169 %	0.931	0.919	0.922
KNN	0.01	66.3657 %	0.891	0.664	0.675
NBM	0.06	92.0638 %	0.927	0.921	0.921

TABLE VII. LIGHT STEMMER EXPERIMENTS BY TAKING NO FEATURE SELECTION TYPE

Classifier type	Time to build model	Null Feature Selection Type			
		Accuracy	Precision	Recall	F-Measures
D.T.	44.05	99.5171 %	0.995	0.995	0.995
NB	6	91.9169 %	0.931	0.919	0.922
KNN	0	66.3657 %	0.891	0.664	0.675
NBM	0.07	92.0638 %	0.927	0.921	0.921

TABLE VIII. NULL STEMMER EXPERIMENTS BY TAKING CHI-SQUARE FEATURE SELECTION TYPE

Classifier type	Time to build model	Chi-square Feature Selection Type			
		Accuracy	Precision	Recall	F-Measures
D.T.	100.98	99.6221 %	0.996	0.996	0.996
NB	16.29	91.329 %	0.923	0.913	0.914
KNN	0.01	66.3867 %	0.781	0.664	0.63
NBM	0.05	92.0638 %	0.928	0.921	0.921

TABLE IX. NULL STEMMER EXPERIMENTS BY TAKING INFO GAIN RATIO FEATURE SELECTION TYPE

Classifier type	Time to build model	Info gain Feature Selection Type			
		Accuracy	Precision	Recall	F-Measures
D,T.	100.5	99.6221 %	0.996	0.996	0.996
NB	17.13	91.329 %	0.923	0.913	0.914
KNN	0	75.0577 %	0.802	0.751	0.734
NBM	0.13	92.0638 %	0.928	0.921	0.921

TABLE X. NULL STEMMER EXPERIMENTS BY TAKING NO FEATURE SELECTION TYPE

Classifier type	Time to build model	Null Feature Selection Type			
		Accuracy	Precision	Recall	F-Measures
D.T.	100.56	99.5801 %	0.996	0.996	0.996
NB	17.54	91.329 %	0.923	0.913	0.914
KNN	0	66.3867 %	0.781	0.664	0.63
NBM	0.2	92.0638 %	0.928	0.921	0.921

V. CONCLUSION

we have been investigated the performance of two FS methods with four classifiers(decision tree, K-nearest neighbors(KNN), Naïve Bayesian method and Naïve Bayes multinomial classifier) using Arabic dataset. The accuracy for decision tree, Naïve Bayesian method and Naïve Bayes multinomial better than K-nearest neighbors(KNN) in all cases. In Future work we will use more feature selection methods with different classifiers algorithms.

REFERENCES

[1] M. Abdelwad. "Chi Square Feature Extraction Based Svms Arabic Language Text Categorization System." *Journal of Computer Science*, 2007.

[2] M.Abdelwadood. "Support vector machines based Arabic language text classification system: feature selection comparative study." Advances in Computer and Information Sciences and Engineering.Springer Netherlands, 2008.

[3] C. Hamoudaand W. David. "Feature subset selection for Arabic document categorization using BPSO-KNN." Nature and Biologically Inspired Computing (NaBIC), Third World Congress on. IEEE, 2011.

[4] H. Bilal, A. Mansour, and Sh. Aljawarneh. "An Efficient Feature Selection Method for Arabic Text Classification." International Journal of Computer Applications ,2013.

[5] M.Abdulrahman, I. Hmeidi, and I. Alsmadi. "Indexing of Arabic documents automatically based on lexical analysis". arXiv preprint arXiv:1205.1602 2012.

[6] A. Saleh. "Automated Arabic Text Categorization Using SVM and NB."Int. Arab J. e-Technology 124-128. 2011.

[7] G, Sami, and N. Ben Amara. "Neural Networks and Support Vector Machines Classifiers for Writer Identification Using Arabic Script." International Arab Journal of Information Technology (IAJIT) 2008.

[8] R. Saleh, et al. "Bilingual experiments with an Arabic-English corpus for opinion mining. 2011.

[9] B, AlSalemi, , and M. Ab Aziz. "Statistical Bayesian Learning for Automatic Arabic Text Categorization." Journal of Computer Science 2011.

[10] A. Ahmed, H. Chen, and A. Salem. "Sentiment analysis in multiple languages: Feature selection for opinion classification in Web forums."ACM Transactions on Information Systems (TOIS)2008.

Implementation of ADS Linked List Via Smart Pointers

Ivaylo Donchev, Emilia Todorova
Department of Information Technologies, Faculty of Mathematics and Informatics
St Cyril and St Methodius University of Veliko Turnovo
Veliko Turnovo, Bulgaria

***Abstract*—Students traditionally have difficulties in implementing abstract data structures (ADS) in C++. To a large extent, these difficulties are due to language complexity in terms of memory management with raw pointers – the programmer must take care of too many details to provide reliable, efficient and secure implementation. Since all these technical details distract students from the essence of the studied algorithms, we decided to use in the course in DSA (Data Structures and Algorithms) an automated resource management, provided by the C++ standard ISO/IEC 14882:2011. In this work we share experience of using smart pointers to implement linked lists and discuss pedagogical aspects and effectiveness of the new classes, compared to the traditional library containers and implementation via built-in pointers.**

Keywords—abstract data structures; C++; smart pointers; teaching

I. Introduction

From the C language we know that pointers are important but are a source of trouble. One reason to use pointers is to have reference semantics outside the usual boundaries of scope [1]. However, it can be quite difficult to ensure that the life of the pointer and the life of the object to which it points will coincide, especially in cases where multiple pointers point to the same object. Such a situation we have as if an object must participate in multiple collections – each of them must provide a pointer to this object. To make everything correct we need to ensure:

- When destroying one of the pointers, take care that there are no dangling pointers or multiple deletions of the pointed object;
- When you destroying the last reference to an object, to destroy the very object in order not to allow resource leaks;
- Do not allow null-pointer dereference – a situation in which a null pointer is used as if it points to a real object.

We must have in mind such details if we want to accomplish dynamic implementation of ADS and often the time for this exceeds the time remaining to comment the structures and operations on them. Moreover, there are rare cases where we have a working implementation of a structure with carefully designed interface and methods written according to the best methodologies, but we identify gaps in the management of memory only when the fall in non-trivial situations such as copying large structures, transfer of items from one structure to another, or destruction of a large recursive structure. For each class representing ADS the programmer must also provide characteristic operations as well as correctly working copy and move semantics, exception handling, construction and destruction. This requires both time and expertise in programming at a lower level. The teacher will have to choose between emphasizing on language-specific features and quality of implementation or to compromise with them and to spend more time on algorithms and data structures. In an attempt to escape from this compromise, we decided to change the content of our CS2 course in DSA, and include the study of smart pointers for resource management and with their help to simplify implementations of ADS, and avoid explicit memory management which is widely recognized as error-prone [2].

Our initial hypothesis was that a correct and effective implementation is possible, which could relieve our work in two directions:

- Operations with whole structures: not having to write destructors, copy and move constructors and copy and move assignment operators;
- Shorter and easier to understand implementation of operations with elements of structures – include (insert element), search, delete.

II. Development of Language Tools for Dynamic Memory Management

Before introducing of `new` and `delete` for work with dynamic memory, inherited from the C language functions `malloc, calloc, realloc` and `free` are used, which are still available in C++ by including the header file `<cstdlib>`.

```
Data * d = (Data *)  malloc(sizeof Data);
// ...
free(d);
```

Memory blocks allocated through these functions are not necessarily compatible with those returned by `new`, so each must be handled with its own set of functions or operations. The problems here are related to unnecessary type conversions and error-prone size calculations (with `sizeof`).

Introduction of `new` and `delete` operators simplifies the syntax, but does not solve all problems. Especially in applications that manipulate complicated linked data structures, it may be difficult to identify the last use of an object. Mistakes

lead to either duplicate de-allocations and possible security holes, or memory leaks [2]. We illustrate this with an example: Let p1, p2, p3 and p4 are pointers to objects of the class Person.

```
vector<Person*> family { p1, p2, p3, p4 };
vector<Person*> kids { p3, p4 };
//...
delete p3;
print(family);//family contains dangling ptr
if (kids.empty()) return 0;     //early return
//...
delete p1;
delete p2;
delete p3;        // double deletion
delete p4;
```

The two vectors – family and kids contain pointers to shared objects – p3 and p4. Deleting the object pointed to by p3 leads to the emergence of "dangling" pointers in the two vectors because they cannot "understand" that the referred object is deleted. All the potential problems with locally defined naked pointers include:

- **Leaked objects**: memory allocation with new can cause (though rarely) an exception which is not handled. It is also possible function execution to be terminated by another raised exception and the allocated with new memory to remain unreleased (it is not exceptions safety). Avoiding such resource leak usually requires that a function catches all exceptions. To handle the deletion of the object properly in case of an exception, the code becomes complicated and cluttered. This is a bad programming style and should be avoided because it is also error prone. Similar situation we have when function execution is terminated by premature return statement based on some condition (early return);
- **Premature deletion**: we delete an object that has some other pointer to and later use that other pointer.
- **Double deletion**: we are not insured against an attempt to re-delete an object (in the example with vectors the one pointed by p3).

One way to circumvent these problems is to simply use a local variable, instead of a pointer, but if we insist to use pointer semantics, the usual approach to overcome such problems is the use of "smart pointers". Their "intelligence" is expressed in that they "know" whether they are the last reference to the object and use this knowledge to destroy the object only when its "ultimate owner" is to be destroyed. We can consider that a "smart pointer" is RAII (Resource Acquisition Is Initialization) modeled class that manages dynamically allocated memory. It provides the same interfaces that ordinary pointers do (*, ->). During its construction it acquires ownership of a dynamic object in memory and deallocates that memory when goes out of scope. In this way, the programmer does not need to care himself for the management of dynamic memory.

For the first time the standard C++98 introduces a single type of smart pointer – auto_ptr which provides specific and focused transfer-of-ownership semantics. auto_ptr is most charitably characterized as a valiant attempt to create a unique_ptr before C++ had move semantics. auto_ptr is now deprecated, and should not be used in new code. It works well in trivial situations:

```
int main(){
   try {
          auto_ptr<X> ap1(new X(1122));
          // _div() throws exception
          cout << _div(5, 0) << endl;
          ap1->print();
   }
   catch (exception& e){
          cerr << e.what() << endl;
   }
}
```

Template auto_ptr holds a pointer to an object obtained via new and deletes that object when it itself is destroyed (such as when leaving block scope). Function _div() returns the quotient of its arguments and causes an exception at zero divisor. Thus, in main() an exception occurs and the operator ap1->print() will not be executed, but still the memory that ap1 manages will be properly released. This is due to the stack unwinding, which occurs in exception processing – all local objects defined in the try block are destroyed, the destruction of ap1 releases the associated memory for the object of class X. Here auto_ptr is "smart" enough, but it appears that the problems entailed outweigh the benefit from it:

- copying and assignment among smart pointers transfers ownership of the manipulated object as well. That is, by default move assignment and move construction is carried out. Such is the situation with passing of auto_ptr as a parameter of the function:

```
void foo(auto_ptr<X> ap2){
   ap2->print();
}
int main(){
   auto_ptr<X> ap1(new X(1122));
   foo(ap1);
   ap1->print();  //oops! ap1 is empty
}
```

After completion of foo() the memory allocated in the initialization of ap1 and then passed to ap2 will be released (at the destruction of ap2) and will not be given back to ap1. This will result in an error when trying to use the contents of ap1 (it is already a dangling pointer).

We have a similar result in the following situations:

```
auto_ptr<X> ap3(ap1);     //move construction
ap1->print();             //oops! ap1 is empty
auto_ptr<X> ap4;
ap4 = ap3;                //move assignment
ap3->print();             //oops! ap3 is empty
```

In constructing ap3 it acquires the resource managed by ap1. This is called *copy elision*. In some cases this is a very useful technique (eg to avoid unnecessary copying when the function returns local object by value – compilers do this automatically).

The auto_ptr provides a semantics of strict ownership. auto_ptr owns the object it holds a pointer to. Copying an auto_ptr copies the pointer and transfers ownership to the destination. If more than one auto_ptr owns the same object at the same time, the behavior of the program is undefined.

- auto_ptr can not be used for an array of objects. When auto_ptr goes out of scope, delete runs on its associated memory block. This works if we have a single object, not an array of objects that must be destroyed with delete [].
- because auto_ptr does not provide shared-ownership semantics, it can not even be used with Standard Library containers like vector, list, map.

Although auto_ptr is now officially deprecated by the standard ISO/IEC, 2011 [4], in Visual Studio 2013 can have declarations like:

```
auto_ptr<vector<int>> apv {new vector<int>{ 1
    2, 3, 4, 5 } };
vector<auto_ptr<int>> v;
```

The reason for this is the famous backward compatibility feature of C ++.

Practice shows that to overcome (or at least limit) problems as described above it is not sufficient to use only one "smart pointer" class. Smart pointers can be smart in some aspects and carry out various priorities, as they have to pay the price for such intelligence [1], p. 76. Note that even now, with several types of smart pointers their misuse is possible and programming of wrong behavior.

In the standard (ISO/IEC, 2011) instead of auto_ptr several different types of smart pointers are introduced (also called Resource Management Pointers) [5], modeling different aspects of resource management. The idea is not new – it formally originates from [6] and was originally implemented in the Boost library and only in 2011 became a part of the Standard Library. The basic, top-level and general-purpose smart pointers are unique_ptr and shared_ptr. They are defined in the header the file <memory>.

Unfortunately, excessive use of new (and pointers and references) seems to be an escalating problem. However, when you really need pointer semantics, unique_ptr is a very lightweight mechanism, with no additional costs compared to the correct use of built-in pointer [5], p. 113. The class unique_ptr is designed for pointers that implement the idea of exclusive (strict) ownership, what was intended auto_ptr to do. It ensures that at any given time only one smart pointer may point to the object. As a result, an object gets destroyed automatically when its unique_ptr gets destroyed. However, transfer of ownership is permitted. This class is particularly useful for avoiding leak of resources such as missed delete calls for dynamic objects or when exception occurs while an object is being created. It has much the same interface as an ordinary pointer. Operator * dereferences the object to which it points, whereas operator -> provides access to a member if the object is an instance of a class or a structure. Unlike ordinary pointers, smart pointer arithmetic is not possible, but specialists consider this an advantage, because it is known that pointer arithmetic is a source of trouble. unique_ptr uses include passing free-store allocated objects in and out of functions (rely on move semantics to make return simple and efficient):

```
// make Person object and give it to a
   unique_ptr
unique_ptr<Person> make_Person(
   const string & name, int year)
{
   // ... check Person, etc. ...
   return unique_ptr<Person>{new Person{name,
   year}};
}
// ..................
auto pp = make_Person("Ivaylo", 1971);
pp->print();
```

For such situations std::move() will be automatically executed for the return value (under the new rules in C++11). Copying or assignment between unique pointers is impossible if we use the ordinary copy semantics. However, we can use the move semantics. In that case, the constructor or assignment operator transfers the ownership to another unique pointer.

The typical use of unique_ptr includes:

- ensuring safe use of dynamically allocated memory through the mechanism of exceptions (exception safety);
- transfer of ownership of dynamically allocated memory to function (via parameter);
- returning dynamically allocated memory – the function returns a pointer to the allocated memory (unique_ptr) ;
- storing pointers in a container.

A point of interest is the situation when unique_ptr is passed as a parameter of a function by rvalue reference, created by std::move(). In this case the parameter of the called function acquires ownership of unique_ptr. If then this function does not pass ownership again, the object will be destroyed at its completion:

```
template <typename T>
void f(unique_ptr<T> x)
{
   cout << *x << endl;
}
int main()
{
 unique_ptr<string> up{new string{"Ivaylo"}};
 f(move(up));     // up became empty
 if (up) cout << *up << endl;
   else cout << "empty pointer" << endl;
}
```

Using a unique pointer, as a member of a class may also be useful for avoiding leak of resources. By using unique_ptr, instead of built-in pointer there is no need of a destructor because the object will be destroyed while destroying the member concerned. In addition unique_ptr prevents leak of resources in case of exceptions which occur during initialization of objects – we know that destructors are called

only if any construction has been completed. So, if an exception occurs within the constructor, destructors will be executed for objects that have been already fully constructed. As a result we can get outflow of resources for classes with multiple raw pointers, if the first construction with new is successful, but the second fails.

Simultaneous access to an object from different points in the program can be provided through ordinary pointers and references, but we already commented on the problems associated with their use. Often we have to make sure that when the last reference to an object is deleted, the object itself will be destroyed as well (which usually implies garbage collection operations – to deallocate memory and other resources).

The class `shared_ptr` implements the concept of shared ownership. Many smart pointers can point to the same object, and the object and its associated resources are released when the last reference is destroyed. The last owner is responsible for the destroying. To perform this task in more complex scenarios auxiliary classes `weak_ptr`, `bad_weak_ptr`, `enable_shared_from_this` are provided.

The class `shared_ptr` is similar to a pointer with counter of the number of sharings (reference counter), which destroys the pointed object when this counter becomes zero. Imagine `shared_ptr` as a structure of two pointers – one to the object and one to the counter of sharings.

Shared pointer can be used as an ordinary pointer – to assign, copy and compare, to have access to the pointed object via the operations `*` and `->`. We have a full range of copy and move constructions and assignments. Comparison operations are applied to stored pointers (usually the address of the owned object or `nullptr` if none). `shared_ptr` does not provide index operation. For `unique_ptr` a partial specialization for arrays is available that provides `[]` operator, along with `*` and `->`. This is due to the fact that `unique_ptr` is optimized for efficiency and flexibility. Access to the elements of the owned by `shared_ptr` array can be provided through the indices of the internal stored pointer, encapsulated by `shared_ptr` (and accessible through the member function `get()`).

We already discussed the problems with dangling pointers, which arise while build-in pointers are stored in containers. Now we will show how the use of `shared_ptr` avoids them. Consider the same situation with vectors of `Person` objects – `family` and `kids`:

In the function `main()` we have 4 shared pointers, to manipulative dynamic objects of `Person`:

```
auto sp1=make_shared<Person>("Ivaylo", 1971);
auto sp2=make_shared<Person>("Doroteya", 1977);
auto sp3=make_shared<Person>("Victoria", 2002);
auto sp4=make_shared<Person>("Peter", 2009);
```

and two vectors of such pointers in which objects are duplicated:

```
vector<shared_ptr<Person>> sp_family{sp1,
   sp2, sp3, sp4};
vector<shared_ptr<Person>> sp_kids{sp3, sp4};
```

There is a single copy of each object of `Person`. The number of references to the children is 3 - one in each vector and the one of `sp3` (or `sp4`).

The name change

```
sp3->set_name("Victoria Doncheva");
```

immediately affects both vectors. Release of `sp3` by `reset()` does not lead to destruction of the object `Person {"Victoria", 2002}`, in opposit to build-in pointers.

Of course, if you like, you can always make a mess. If you initialize a build-in pointer with the owned by `shared_ptr` internal pointer, and then deallocate memory by this raw pointer:

```
Person* p = sp3.get();
delete p;
```

A problem with reference-counted smart pointers is that if there is a ring, or cycle, of objects that have smart pointers to each other, they keep each other "alive" – they will not get deleted even if no other objects are pointing to them from "outside" the ring. Such a situation often occurs in implementations of recursive data structures. C++11 includes a solution: "weak" smart pointers: these only "observe" an object but do not influence its lifetime. A ring of objects can point to each other with `weak_ptr`s, which point to the managed object but do not keep it in existence. Like raw pointers, the weak pointers do not keep the pointed-to object "alive". The cycle problem is solved. However, unlike raw pointers, the weak pointers "know" whether the pointed-to object is still there or not and can be interrogated about it, making them much more useful than a simple raw pointer would be.

In practice often happens a situation when we hesitate which version of a smart pointer to use – `unique_ptr` or `shared_ptr`. The advice is to prefer `unique_ptr` by default, and we can always later move-convert to `shared_ptr` if needed. There are three main reasons for this [7]:

- try to use the simplest semantics that are sufficient;
- a `unique_ptr` is more efficient than a `shared_ptr`. A `unique_ptr` does not need to maintain reference count information and a control block under the covers, and is designed to be just about as cheap to move and use as a raw pointer;
- starting with `unique_ptr` is more flexible and keeps your options open.

In our case, however, we had from the very beginning to start with `shared_ptr`, because being recursive by definition, the data structures that we tried to implement with smart pointers can not do without shared ownership.

III. Implementation of Lists

In the course in Data Structures and Algorithms (DSA) we use dynamically implemented singly linked and doubly linked lists and based on them specializations for other ADS – stack, queue, deque. We develop a template class `List` with an interface similar to the following:

```
//singly linked list with built-in pointers
```

```
template <typename T>
class List {
private:
   struct Node {
      T key;
      Node* next;
      Node():key(),next(nullptr){}
      Node(T x):key(x),next(nullptr){}
   };
   Node* front;       //first element
public:
   List():front(nullptr){}   //default constructor
   List(T x):front(new Node(x)){}
   //initializer list constructor
   List(initializer_list<T>);
   ~List();           //destructor
   List(const List&); //copy constructor
   List(List&&);      //move constructor
   //copy assignment
   List& operator =(const List&);
   List& operator =(List&&); //move assignment
   bool push_front(T);       //add to the top
   bool push_back(T);        //add to the bottom
   T& operator [](int);      //index operator
   size_t size();     //the length of the list
   bool find(T);      //search for element;
   Node* find_ref(T); //reference to element
   bool empty(){ return front == nullptr; }
   bool remove(T);
};
```

In addition, students develop on their own methods to insert a node in any location; to search and insert an element in a way to keep the list sorted; to exchange places of elements; to insert an element before and after a node; to merge two lists and more.

Since we count on the reliability, in the course we try to follow the methodology for verification of object-oriented programs as proposed in [3]. Correct implementation of all methods requires multiple checks; catching any exceptions; tracking the number of references to a node. Our current practice shows that students encounter the greatest difficulties in removing items from the list and the most common mistake is to forget a `delete` operator in any branch of the algorithm. So in fact an element is excluded from the list, but the occupied memory is not released – a typical example of a memory leak. Other typical logic errors are skipping a special case such as an attempt to delete an item from an empty list or when the element to be deleted is the first in the list.

In order to simplify the technical part and to focus on algorithms, implementing the operations on lists from 2013-2014, we went to implementation with smart pointers. Our initial expectation was that it was possible to avoid all methods of copy and move semantics, destructors for nodes and list, release of memory when deleting nodes and exception handling related to the construction of a list and its nodes. We relied on simplified syntax in the implementation of operations.

We started with the realization of the template class with the following interface:

```
template <typename T>
class List {
    class  Node {
    public:
        T key;
        shared_ptr<Node> next;
        Node():key(), next(){}
        Node(T x):key(x), next(){}
    };
    shared_ptr<Node> top;
    shared_ptr<Node> bottom;
public:
    List():top(), bottom(make_shared<Node>()){}
    List(T x):top(make_shared<Node>(x)),
              bottom(make_shared<Node>()){
       top->next = bottom;}
    List(initializer_list<T>);
    bool push_front(T);
    bool push_back(T);
    operator bool(){return top!=nullptr;}
    shared_ptr<Node> find(T)
    bool remove(const T&);
    T& operator[](size_t);
};
```

Unlike the interface of `std::forward_list`, we added a feature inserting elements at the end (the method `push_back`) and aiming a more effective implementation of this, we used a fictitious node `bottom` as a sentinel.

We will show the advantage of using shared pointers through the method `remove` to delete element with a key `x`:

```
template <typename T>
bool List<T>::remove(const T& x) {
   if(!top) return false;
   if(top->key == x) {
          top = top->next;
          return true;
   }
   for(auto p=top; p->next; p=p->next)
          if(p->next->key == x) {
                 if(p->next == bottom)
                 bottom = p;
                 p->next = p->next->next;
          }
   return true;
}
```

It is seen that the code with shared pointers differs from that with build-in pointers only by avoiding `delete` several times to release occupied by the deleted node memory. The code of the other methods is sufficiently clear and concise, for example adding a new element to the beginning of the list looks like this:

```
template <typename T>
bool List<T>::push_front(T x) {
   auto p = make_shared<Node>(x);
   if(!p) return false;
   p->next = top ? top : bottom;
   top = p;
   return true;
}
```

With automatic type deduction and factory function `make_shared` (row 2) we even avoid explicit type declaration for smart pointer `p` and do not use `new`, instead:

```
shared_ptr<Node> p { new Node{ x } };
```

For educational purposes all operations with a single list ran normally, but when we tested a larger list (100000 strings), we got a "stack overflow" error during the automatic destruction of the list at the end of the program. Because of the recursive links a situation occurs where one node keeps "alive" the whole structure. This on one hand requires a large stack, and on the other – can lead to significant delays in the demolition of the structure. So we decided to add a destructor, instead of increasing the stack size from the settings of the linker:

```
template<typename T>
List<T>::~List() {
   while (top != bottom)
        top = top->next;
}
```

Here, again, we don't use `delete` to release the memory occupied by each node, but instead just sequentially shift the first element until we reach the end of the list. This causes automatic execution of a destructor for each node, managed by shared pointer, as there will be no more references to it.

Further, when working with two or more lists, we encountered problems with copy assignment and copy construction. Both operations performed shallow copying and we had to add a copy constructor and copy assignment operator to evoke correct actions for deep copying. Their code proved to be with complexity equivalent to the version with naked pointers, so in this case we could not save the students the technical details.

The situation with move semantics proved to be analogous – the lack of user-defined move constructor and move-assignment operator results in that after the transfer of ownership the pointer members of the object (list) on the right are not reset to its initial state, so we implemented these methods as well, but as seen from the code below, the implementation is quite trivial and does not burden the students:

```
template<typename T>
List<T>::List(List<T>&&other):
   top(move(other.top)),
   bottom(move(other.bottom)) {
          other.top = nullptr;
          other.bottom = nullptr;
   }
```

The reason that compiler-generated move semantics methods don't work is that the complex types, such as our list, often define one or more of the special member functions themselves, and this can prevent other special member functions from being automatically generated. This problem we solved in another way, without implementation of the corresponding methods, but passed to compiler that supports explicitly defaulted and deleted functions – Microsoft Visual C++ Compiler Nov 2013 CTP (CTP_Nov2013).

Then the declarations

```
List(List&&) = default;
List& operator =(List&&) = default;
```

provided smooth operation of the automatically generated move constructor and move-assignment operator. Unfortunately we found that this approach does not work with copy semantics.

Similar difficulties were encountered with the implementation of Doubly Linked List. Here is a part of its interface:

```
template <typename T> class List {
    class Node {
    public:
           T key;
           shared_ptr<Node> next;
           weak_ptr<Node> prev;
           Node():key(),next(), prev(){}
           Node(T x):key(x), next(), prev(){}
    };
    shared_ptr<Node> front;
    shared_ptr<Node> back;
public:
    List():front(), back(){}
    List(initializer_list<T>);
    bool push_front(T);
    bool push_back(T);
//...
};
```

As here are bidirectional links, in order not to duplicate them and make the structure "indestructible", for those in the opposite direction we use a weak pointer. And for this list we can state that implementation of operations has the same or less complexity than the version with built-in pointers.

We will comment on another issue, connected not so much with the lists as with the syntax rules in C++11 and implementation of initializer list constructor. If you try to initialize a list with another using the syntax for uniform initialization:

```
List <int> L2 {L1};
```

If we have templatized initializer list constructor, the compiler will consider this as a call to this constructor with an argument initializating list of one element of type `List<int>`, not as a call to the copy constructor. That would cause unexpected behavior. One option for dealing with the problem is definition of specialization for initializer list constructor for lists:

```
List(initializer_list<List<T>>);
```

The other option is simply to use function syntax:

```
List<int> L2(L1);
```

In conclusion we can assert that although our initial idea to avoid implementation of all special member functions was not completely accomplished, these methods, as well as all operations with lists can be implemented more concisely and clearly than their respective analogues in the build-in pointers implementation. Furthermore, by using smart pointers we implemented a complete "no naked new " policy, respecting the recommendation of [5], p. 64 that avoiding naked `new` and naked `delete` makes code far less error-prone and far easier to

keep free of resource leaks. From this perspective, we consider reasonable study of smart pointers in the course of DSA.

IV. Performance Evaluation

In order to evaluate the efficiency of smart pointers implementation we carried out an experiment in which we compare the times for typical operations with lists, implemented with and without smart pointers.

Three implementations of Singly Linked Lists with library `std::forward_list` and our realization of Doubly Linked List with smart pointers with library equivalent `std::list` we compared (Table 1). The same data is used in the experiment: 100'000 randomly generated unique strings of length of 20 stored in a text file. They are used to construct lists by adding elements to the beginning for the one-directional linked versions and at the end of bi-directional linked lists.

TABLE I. Test Results

Operations	**List Implemantations**					
	Singly Linked Lists				***Doubly Linked Lists***	
	C-style	Row Pointers	Smart Pointers	std::forward_list	Smart Pointers	std::list
Add node	78	109	109	78	125	63
Traverse	14 078	14 703	30 578	19 829	31 829	14 546
Delete node	21 594	21 625	143 703	107 515	153 812	78 172

Note: Time in milliseconds

The first operation "Add element" reads all strings from the file and stores them in the relevant list. For each list the text file is opened and read again.

Traversing accomplishes 10000 searches for an element not contained in the list: that is complete pass over all the nodes.

The test of deletion is deliberately made so as to require multiple traverse – check if each element meets the set criterion (comparison of strings) and if so, the key of this element is passed as argument to the deleting function. This function each time searches the element from the beginning of the list and deletes only the first hit. 59 996 elements of all 100 000 are deleted.

The results show a negligibly small difference in performance between the implementation without classes (C-style), and implementation using classes and raw pointers. Only the "add element" operation is 28% slower. Time difference between single linked lists and bi-directional linked lists implemented with smart pointers is inessential. This was expected because the test algorithms traverse lists only in one direction. The advantage of bi-directional linked list is only visible in comparison with library implementations. The library template class `forward_list` is inferior in efficiency to our raw pointer implementation for traverse operation by 26%, and removing elements is nearly 5 times slower. Implementation of smart pointers has significantly weaker results – traverse is 2 times slower, and removing elements – 6 times compared to raw pointers. Adding elements shows no difference in performances. Our version of bidirectional linked list with smart pointers proved to be twice slower than library version `std::list` for all operations.

V. Conclusion

Our initial hypothesis regarding the implementation of lists with smart pointers was proven partially. We could not do entirely without implementation of methods of copy and move semantics, but their code turned out to be short, clear and easily understandable for students. Moreover, move semantics in our case can be provided by defaulted move constructors and assignment operators. We consider the second part of the hypothesis, namely the shorter and clearer implementation of the basic operations with data structures for fully achieved. In addition, smart pointer versions do not require user-defined exception handling.

Since we do not have enough empirical data, we cannot prove the advantage of this way of teaching DSA yet, but even without conducting a strictly formal pedagogical experiment, we can confirm that the results of students tests, homework and exams are comparable to those demonstrated by their colleagues trained in previous years under the old program.

The implementation of ADS with smart pointers is more clear and concise, but requires spending time to study in additionally templates and essential elements of the STL, though not in detail. This could be facilitated by reorganizing CS1 course Programming Fundamentals, where to underlie learning C++11/14 and STL. Note that for our implementations it is not needed even to know the full interface for work with smart pointers. In most situations the interface of build-in pointers is sufficient plus function `make_shared` and possibly member function `reset`. In our work with the students during the school year we met difficulties in debugging of programs related to discovery of logical errors in memory management, most often connected with its release. We found that it is appropriate to add an intermediate output (operator `cout`) in the destructors as of DSA, as of the elements held in them (if they are of user-defined types). In this way it is easy to detect situations where objects remain undestroyed.

Regarding the applicability of smart pointers in the actual programming will mention the opinion of Stroustrup, that they "are still conceptually pointers and therefore only my second choice for resource management – after containers and other types that manage their resources at a higher conceptual level" [5], p. 114. The results of our comparative tests also show that library containers are sufficiently effective and can join the opinion of Stroustrup. Furthermore, anyway, to learn smart pointers it is necessary to get into STL. On one hand it is better to teach students how to use its efficient and reliable containers. On the other hand though, we train professionals and they must be able to independently implement such containers – to develop creative thinking. It is therefore not a bad idea to do so with smart pointers as well – one more opportunity provided by the STL.

References

[1] Josuttis, N. M. (2012). *The C++ Standard Library: A Tutorial and Reference*. Addison-Wesley Professional; 2nd edition (April 9, 2012).

[2] Boehm, H. & Spertus, M. (2009). *Garbage Collection in the Next C++ Standard*. Proceedings of the 2009 international symposium on Memory

management, pp. 30-38. ACM New York. doi>10.1145/1542431.1542437

[3] Todorova, M., Kanev, K. (2012). *Educational framework for verification of object-oriented programs*, in Proceedings of the 2012 Joint International Conference on Human-Centered Computer Environments, ACM, New York, pp. 23-27

[4] ISO/IEC. (2011). *International Standard ISO/IEC 14882:2011(E) Information technology – Programming languages – C++* (3rd ed.)

[5] Stroustrup, Bj. (2013). *The C++ Programming Language, 4th Edition*. Addison-Wesley Professional; 4th edition (May 19, 2013)

[6] Dimov, P., Dawes, B. & Colvin, G. (2003). *A Proposal to Add General Purpose Smart Pointers to the Library Technical Report*. C++ Standards Committee Papers. Document number: N1450=03-0033 http://www.open-std.org/jtc1/sc22/wg21/docs/papers/2003/n1450.html

[7] Sutter, H. (2013). *Sutter's Mill. GotW #89 Solution: Smart Pointers*. http://herbsutter.com/2013/05/29/gotw-89-solution-smart-pointers/

A Review on Feature Extraction and Feature Selection for Handwritten Character Recognition

Muhammad 'Arif Mohamad
Faculty of Computing
Universiti Teknologi Malaysia
Johor Bharu,Malaysia.

Dewi Nasien
Faculty of Computing
Universiti Teknologi Malaysia
Johor Bharu,Malaysia.

Haswadi Hassan
Faculty of Computing
Universiti Teknologi Malaysia
Johor Bharu,Malaysia.

Habibollah Haron
Faculty of Computing
Universiti Teknologi Malaysia
Johor Bharu,Malaysia.

***Abstract*—The development of handwriting character recognition (HCR) is an interesting area in pattern recognition. HCR system consists of a number of stages which are preprocessing, feature extraction, classification and followed by the actual recognition. It is generally agreed that one of the main factors influencing performance in HCR is the selection of an appropriate set of features for representing input samples. This paper provides a review of these advances. In a HCR, the set of features plays as main issues, as procedure in choosing the relevant feature that yields minimum classification error. To overcome these issues and maximize classification performance, many techniques have been proposed for reducing the dimensionality of the feature space in which data have to be processed. These techniques, generally denoted as feature reduction, may be divided in two main categories, called feature extraction and feature selection. A large number of research papers and reports have already been published on this topic. In this paper we provide an overview of some of the methods and approach of feature extraction and selection. Throughout this paper, we apply the investigation and analyzation of feature extraction and selection approaches in order to obtain the current trend. Throughout this paper also, the review of metaheuristic harmony search algorithm (HSA) has provide.**

Keywords—HCR; Feature Extraction; Feature Selection; Harmony Search Algorithm

I. INTRODUCTION

Handwriting Character Recognition (HCR) is the ability of a computer to receive and interpret intelligible handwritten input then analyzed to many automated process system. Generally, HCR can be divided into three steps namely pre-processing, feature extraction and classification (recognition). Preprocessing stage is to produce a clean character image that can be used directly and efficiently by the feature extraction stage. Feature extraction stage is to remove redundancy from data. Classification stage is to recognize characters or words. This paper only concentrates in the feature extraction stage.

HCR is a challenging problem since there is a variation of same character due to the change of fonts and sizes. The differences in font types and sizes make the recognition task difficult and resulting the recognition of character process becomes not good.

Feature extraction in HCR is a very important field of image processing and object recognition. Fundamental component of characters are called features. The basic task of feature extraction and selection is to find out a group of the most effective features for classification; that is, compressing from high-dimensional feature space to low-dimensional feature space, so as to design classifier effectively [1].

Based on the statement above, this study was conducted to review and examine the approach as extraction and selection method for feature in HCR. This study also was conducted to investigate a current trend on approach of feature extraction and selection.

This paper is divided to four sections. Section I describes introduction. Section II describes overview on HCR. Section III describes overview on feature extraction followed by current trend on feature extraction in next section IV. For Section V and VI, an overview on feature selection and current trend on feature selection were describes respectively. In Section VII the discussion and future work were discussed briefly. The last section shows conclusion of the whole content.

II. EASE OF USE

The Handwriting recognition is defined as the transformation of a language into symbolic representation from its visual marks [3]. The goal of handwriting recognition is to interpret input where it can be recognition of handwritten sentences, words or characters. Character recognition is a part of a handwriting recognition problem. The development of handwriting character recognition (HCR) is an interesting area in pattern recognition or sometimes specifically referred as optical character recognition (OCR),

According to Arica and Yarman-Vural in their review of character recognition (CR), the CR systems have evolved in three stages [4]. The early stage is in the period of 1900-1980. The beginning of OCR was said to have started with the objective of developing reading machines for the blind.

In these early systems of automatic recognition of characters, area of concentrations are either in machine printed text or upon small sets of well-distinguished handwritten text or symbols. In the second period of development in the era of the 1980s to 1990s, the explosion of information technology has helped a rapid growth in the area of OCR. The CR research was focused basically on the shape recognition techniques without using any semantic information. Although an upper limit in the recognition rate was achieved, it was not sufficient in many practical applications. The 1990s and onwards are referred as the advancements era [5], where the real progress in OCR systems has been achieved. In the beginning of this period, image processing and pattern recognition techniques were efficiently combined with artificial intelligence (AI) methodologies. Complex algorithms for character recognition systems were developed. There is, however, still a long way to go in order to reach the ultimate goal of machine simulation of fluent human reading, especially for unconstrained on-line and off-line handwriting [4].

HCR can be divided into two categories namely, online and off-line. On-line character recognition involves the identification of characters while they are written [6] and deals with time ordered sequences of data, pen up, and down movement and pressure sensitive pads that record the pen's pressure and velocity [7]. On the other hand, off-line character recognition involves the recognition of already written character patterns in scanned digital image. The off-line character recognition is more complex and requires more research compared to on-line character recognition.

HCR is a very complex task since different writing styles and handwriting variability can produce extreme differences in characters [8,9]. The handwriting development is more sophisticated, found in various kinds of handwritten character such as digit, numeral cursive script, including English, Tamil, Chinese, Bangla, Devanagari, Persian, Arabic and others. The problem and difficulties of handwriting recognition task can be classified into four categories which are nature of the handwriting signals, handwriting styles, writer dependency and vocabulary sizes [10].

Most current approaches to HCR which consist of three main stages namely pre-processing, feature extraction and classification.

A. Preprocessing

The preprocessing stage aims to extract the relevant textual parts and prepares them for segmentation and recognition. The main objectives of preprocessing are noise reduction, normalization of data and compression in the amount of information to be retained [11]. In noise reduction alone there are hundreds of available techniques which can be categorized into three major groups of filtering, morphological operations and noise modeling [12][13]. Filters can be designed for smoothing [14], sharpening [15], thresholding [16], removing slightly textured background [17] and contrast adjustment processes [18]. Various morphological operations can be designed to connect broken strokes [19], decompose the connected strokes [20], smooth the contours, prune the wild points, thin the characters [21], and extract boundaries [22]. Preprocessing stage is to produce a clean character image that can be used directly and efficiently by the feature extraction stage.

B. Feature Extraction

Feature extraction stage is to remove redundancy from data. Before building the feature extraction procedure, there are two important problems must be clarified which are feature extraction and feature selection. Feature extraction is related with which technique will be used to extract features from the image character as representations. On the other hand, in feature selection, the most relevant features to improve the classification accuracy must be searched. This paper only concentrates in the feature extraction and selection stage. The next section wills discuss the feature extraction briefly.

C. Classification

Classification stage is to recognize characters or words. After features that represent the raw input data are extracted, classification stage would use the data to recognize the feature class based on the properties in the features. There are many techniques available in the classification method that can be applied. The classification method can be traced from template matching [23-25], statistical approach [26-28], syntactic [29] and neural network [30].

III. AN OVERVIEW ON FEATURE EXTRACTION

Feature extraction can be defined as extracting the most representative information from the raw data, which minimizes the within class pattern variability while enhancing the between class pattern variability. For this purpose, a set of features are extracted for each class that helps distinguish it from other classes, while remaining invariant to characteristic differences within the class [31]. A good survey on feature extraction methods for character recognition can be found in [32].

Generally there are two kinds of features, statistical features and structural features [33-34], [35-36]. Statistical features contain pixel density, moment, mathematical transformation and so on. Structural features conclude stroke, contour, number of bifurcation points, number of circles and so on. Most researchers agree that statistical features could be obtained quickly using easy methods and could perform good recognition results especially in closed testing data, but it could also be easily affected by the deformation of symbols, thus could not be expanded to more applications. Structural features are more conformed to the intuitive thinking of human mind, thus are more robust for the deformation of symbols. But they usually rely on human summarized rules for the recognition algorithm. When new symbols are introduced into an application, they need more cost to revise the algorithm. [37].

A. Statistical Features

Representation of a document image by statistical distribution of points takes care of style variations to some extent. Although this type of representation does not allow the reconstruction of the original image, it is used for reducing the dimension of the feature set providing high speed and low complexity. The major statistical features mentioned below are used for character representation.

- *Zoning*: The frame containing the character is divided into several overlapping or non-overlapping zones. The densities of the points or some features in different regions are analyzed [38].
- *Crossings and Distances*: A popular statistical feature is the number of crossing of a contour by a line segment in a specified direction. The character frame is partitioned into a set of regions in various directions and then features of each region are extracted.
- *Projections*: Characters can be represented by projecting the pixel gray values onto lines in various directions. This representation creates one-dimensional signal from a two dimensional image, which can be used to represent the character image [39].

B. Structural Features

Structural features are based on topological and geometrical properties of the character. Various global and local properties of characters can be represented by geometrical and topological features with high tolerance to distortions and style variations. This type of representation may also, encode some knowledge about the structure of the object or may provide some knowledge as to what sort of components make up that object. Various topological and geometrical representations can be grouped in four categories:

- *Extracting and Counting Topological Structures*: In this category, lines, curves, splines, extreme points, maxima and minima, cups above and below a threshold, openings, to the right, left, up and down, cross (X) points, branch (T) points, line ends (J), loops (O), direction of a stroke from a special point, inflection between two points, isolated dots, a bend between two points, horizontal curves at top or bottom, straight strokes between two points, ascending, descending and middle strokes and relations among the stroke that make up a character are considered as features [40-41].Avoid combining SI and CGS units, such as current in amperes and magnetic field in oersteds. This often leads to confusion because equations do not balance dimensionally. If you must use mixed units, clearly state the units for each quantity that you use in an equation.
- *Measuring and Approximating the Geometrical Properties*: In this category, the characters are represented by the measurement of the geometrical quantities such as, the ratio between width and height of the bounding box of a character, the relative distance between the last point and the last y-min, the relative horizontal and vertical distances between first and last points, distance between two points, comparative lengths between two strokes, width of a stroke, upper and lower masses of words, word length curvature or change in the curvature[42-45].
- *Coding*: One of the most popular coding schemes is Freeman's chain code. This coding is essentially obtained by mapping the strokes of a character into a 2-dimensional parameter space, which is made up of codes. There are many versions of chain coding. The character frame is divided to left-right sliding window and each region is coded by the chain code [44],[46-48].
- *Graphs and Trees*: Words or characters are first partitioned into a set of topological primitives, such as strokes, holes, cross points etc. Then, these primitives are represented using attributed or relational graphs. Image is represented either by graphs coordinates of the character shape or by an abstract representation with nodes corresponding to the strokes and edges corresponding to the relationships between the strokes. Trees can also be used to represent the words or characters with a set of features, which has a hierarchical relation [49-51].

IV. Current Trends in Feature Extraction

Instead of focusing on feature vector based on a single representation of a character, it is a trend now of combining different types of features extracted from different representations of the same character. The advantage of combining, and harnessing, such different kinds of features is that it can offer wider range of identification clues to help improve the accuracy of recognition. For example, Heutte et al. [52] combine different statistical and structural features for recognition of handwritten characters. They construct a 124-variable feature vector comprising following seven families of features: 1) intersection of the character with horizontal and vertical straight lines, 2) invariant moments, 3) holes and concave arcs, 4) extremas, 5) end points and junction points 6) profiles, and 7) projections. Aurora et al. [53] combine different feature extraction techniques such as intersection based features, shadow features, chain code and curve fitting features for Indian Devnagari language script. Kimura et al. [54] propose a genetic algorithm based strategy for finding a suitable combination of features from a large pool of features with the objective criteria to minimize the classification error. Other combining or hybrid method for features extraction shows in TABLE 1.

TABLE I. Summarization of Combining Feature Extraction Method

Feature Extraction Method	
Author (Year)	***Hybrid Feature Extraction Method***
Vamvakas et al. (2010) [55]	Zoning based features/upper and lower character profile projections features/left and right character profile projections features/ distance based features
Chacko *et al.* (2011) [56]	Wavelet features/chain code features
Wang & Sajjahar (2011) [57]	Polar transformed images/ Zone based feature extraction
Yang *et al.* (2011) [58]	Structural features/Statistical features
Chel *et al.* (2011) [59]	Transition Feature/ Sliding Window Amplitude Feature/ Contour Feature
Al-Khateeb *et al.* (2011) [60]	Structural features/Statistical features
Rajput & Horakeri (2011) [61]	Boundary-based descriptors/namely/ crack codes / Fourier descriptors
Sharma & Jhajj (2011) [62]	Zoning/ Directional Distance Distribution (DDD) / Gabor methods

Feature Extraction Method	
Author (Year)	***Hybrid Feature Extraction Method***
Choudhary *et. al* (2012) [63]	Vertical/ Horizontal/ Left Diagonal and Right Diagonal directions
Li *et al.* (2012) [64]	Direction string / nearest neighbor matching
Nemouchi *et al.* (2012) [65]	Structural(like strokes, concavities, end points, intersections of line segments, loops, stroke relations) / statistic (zoning, invariants moments, Fourier descriptors, Freeman chain code) features
Ahmed *et al.* (2012) [66]	Multi Zoning of the character array (i.e., dividing it into over- lapping or non-overlapping regions, computing the moments of the black pixels of the character, the n-tuples of black or white or joint occurrence, the characteristic loci, and crossing distances)
Likforman-Sulem *et al.* (2012) [67]	Structural and statistic features
Kessentini *et al.* (2012) [68]	Directional density / (black) pixel densities features
Bhattacharya *et al.* (2012) [69]	Chain code computation/ gradient feature / pixel count feature generation
Reddy (2012) [70]	Vertical and horizontal projection profiles (VPP-HPP)/zonal discrete cosine transform (DCT)/ chain-code histograms (CCH) / pixel level values
Muhammad *et al.* (2012) [71]	Correlation based function features/structural/statistical
Vidya V et al. (2013) [72]	Cross feature/ fuzzy depth/distance/ Zernike moment
Primekumar et al. (2013) [73]	Structural Feature/Directional

V. Overview on Feature Selection

In In HCR, feature selection is a technique to select the features that is relevant for classification stage. The goal of feature selection (FS) is that of reducing the number of features to be considered in the classification stage. This task is performed by removing irrelevant or noisy features from the whole set of the available ones. Feature selection is accomplished by reducing as much as possible the information loss due to the feature set reduction: thus, at list in principle, the selection process should not reduce classification performance. The feature selection process consists of three basic steps (see Fig. 1): a search procedure, a subset evaluation and a stopping criterion. A typical search procedure uses a search strategy for finding the optimal solution, according to a given subset evaluation criterion previously chosen. The search procedure is repeated until a stopping criterion is satisfied.

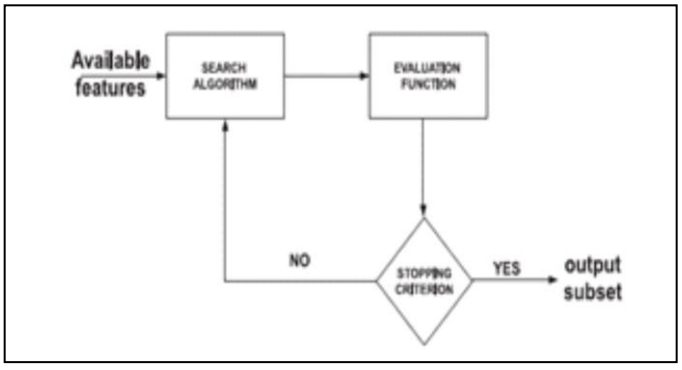


Fig. 1. The feature selection process. [74]

The feature selection problem implies the selection, from the whole set of available features, of the subset allowing the most discriminative power. The choice of a good feature subset is crucial in any classification process because of various reasons. If the considered feature set does not include all the information needed to discriminate samples belonging to different classes, the achievable performance may be unsatisfactory, regardless of the learning algorithm effectiveness. On the other hand, the size of the feature set used to describe the samples determines the search space to be explored during the learning phase. Therefore, irrelevant and noisy features make the search space larger, increasing the complexity of the process. Finally, the computational cost of classification depends on the number of features used to describe the patterns.

When the cardinality N of the candidate feature set Y is high, the problem of finding the optimal feature subset, according to a given evaluation function, becomes computationally intractable because of the resulting exponential growth of the search space, made of all the 2N possible subsets of Y. Therefore, heuristic algorithms become necessary for finding near–optimal solutions [75]. Such algorithms require both the definition of a strategy for selecting feature subspaces and the definition of a function for evaluating the goodness of each selection performed, i.e. how well classes result separated in the selected feature subspace.

As regards evaluation methods, those proposed in the literature can be divided in two wide classes: i) filter methods, which evaluate a feature subset independently of the classifier and are usually based on some statistical measures of distance between the samples belonging to different classes. ii) wrapper methods, which are based on the classification results achieved by a given classifier. Filter methods are usually faster than wrapper ones, as these latter require a new training of the used classifier at each evaluation. Moreover, filter–based evaluations are more general, as they exploit statistical information about the data, while wrapper methods are dependent on the classifier used.

As for search strategies, many heuristic algorithms have been proposed in the literature for finding near–optimal solutions: Greedy selection [76], branch and bound (B&B) [77], floating search [77]. These algorithms use greedy stepwise strategies that incrementally generate feature subsets by adding the feature that produces the highest increment of the evaluation function. Since these algorithms do not take into account complex interactions among several features, in most of the cases they lead to sub–optimal solutions. An alternative way to cope with the search problem is that of using genetic algorithms (GAs), which have demonstrated to be an effective search tools for finding near–optimal solutions in complex and non–linear search spaces [78]. These properties, make GA's suitable also to solve feature selection problems [79-80]. Comparative studies have demonstrated the superiority of GA's in feature selection problems involving large numbers of features [81].

VI. Current Trend in Feature Selection

Recently, interest in feature selection has been on the increase with the abundance of algorithms derived. The feature selection algorithm can be classified into two namely heuristic and metaheuristic approaches. Many heuristic algorithms have

been proposed in the literature for finding near–optimal solutions [76-77].GA is a one of metaheuristic approach and have been widely used to solve feature selection problems [82-87].

TABLE II. SUMMARIZATION OF FEATURE SELECTION APPROACH

Author. (Years)	Feature Selection	
	Approach	*Algorithm*
Nasien D. et al. (2010) [88]	Meta heuristic	Genetic Algorithm (GA) and Ant Colony Optimization (ACO)
Reza A. et al. (2010) [89]	Metaheuristic	Hybrid Genetic Algorithm (GA) + Simulated Annealing (SA)
A. Marcano-Cedeno et al. (2010) [90]	Heuristic	Sequential Forward Selection
Nasien D. et al. (2011) [91]	Heuristic	Randomized and Enumeration based Algorithms
Abandah G. et al. (2011) [92]	Heuristic	Scatter criterion Symmetric uncertainty Fast correlation-based filter (FCBF) Minimal-redundancy-maximal-relevance (mRMR) Non-dominated sorting genetic algorithm (NSGA)
Das N. et al. (2012) [93]	Metaheuristic	Genetic Algorithm based Region Sampling
Roy A. et al. (2012) [94]	Metaheuristic	Artificial Bee Colony (ABC)
Li L. et al. (2012) [64]	Metaheuristic	Nearest Neighbor (NN)
Nagasundara K.B. et al. (2012) [95]	Metaheuristic	Multi cluster feature selection (MCFS)
Stefano et al. (2014) [96]	Meta heuristic	Genetic Algorithm (GA)
Roy A. et al. (2014) [97]	Metaheuristic	Axiomatic Fuzzy Set (AFS)
Ghareh Mohammadi F. et al. (2014) [98]	Metaheuristic	Artificial Bee Colony (ABC)

VII. DISCUSSION AND FUTURE WORK

We have reviewed the introduction, concept and stages in the development of HCR. The goal of handwriting is to identify input characters or image correctly then being analyzed to many automated process system. A handwritten character recognition system consists of a number of preprocessing steps, feature extraction, and classification. This paper only concentrates on feature extraction and selection method.

One of the most important phases in successfully achieving character recognition is the task of feature extraction and selection. Feature extraction is related with which technique will be used to extract features from the image character as representations. On the other hand, in feature selection, the most relevant features to improve the classification accuracy must be searched.

Generally there are two kinds of features extraction, statistical features and structural features. We have investigated and analyzed the method used by researchers to extract the feature for feature extraction. Based on the analysis, recently the most method used by researchers for feature extraction are combining different types of features extracted from different representations of the same character Instead of focusing single representation of a character. It is a trend now of combining different types of features extracted as shown in the Table 1.

On the other hand, the main goal of feature selection is to choose a number of features from the extracted feature set that yields minimum classification error. Meanwhile the feature selection is a technique to select the feature that is relevant to for classification stage. This task is performed by removing irrelevant or noisy features from the whole set of the available ones. Generally, feature selection is finding a subset of features which improve the recognition accuracy. This process has two main phases. First phase includes a search strategy to select one feature subset among all possible, the second phase includes a method for evaluating selected subsets with assigning a fitness value to them generally divided in two: filter and wrapper method.

We have investigated and analyzed the method used by researchers for feature selection. There are many methods or approaches as search strategy for feature selection used by researcher. We were divided in two categories: heuristic and metaheuristic approaches. Based on analysis, recently many researchers used metaheuristic approach rather than heuristic approach for feature selection as shown in TABLE II.

Nowadays, the metaheuristic algorithms have taken an important place in the optimization fields. Metaheuristic algorithm is an approach to solve the optimization problems and to find the best of all possible of solutions. There are many metaheuristic algorithms like genetic algorithm (GA), simulated annealing (SA), particle swarm optimization (PSO) and others. As shown in Table 2, we can see that a researcher used metaheuristic approach i.e. Artificial Bee Colony (ABC) [94], GA and ACO [88] and Axiomatic Fuzzy Set (AFS) [97].

Harmony search algorithm (HSA) is one of the recent metaheuristic that inspired from the musician performance that search for the better state of harmony [98]. To date, HSA has been applied to many engineering optimization problems including structural engineering [99–109], structural materials [110–112], hydraulics [113–116], cost optimization and construction management [117–119], and structural vibration control [120].

As a future work, we considered to propose a feature selection method based on HSA. To the best our knowledge, the HSA have not implemented for feature selection problem yet. Due to the literature study, HS possess several advantages over traditional optimization techniques [121] such as:

- HS is a simple population based metaheuristic algorithm and does not require initial value settings for decision variables;
- HS uses stochastic random searches;
- HS does not need derivation information;
- HS has few parameters;

- HS can be easily adopted in various types of optimization problems [122].

These features increase the flexibility of the HS algorithm in producing better solutions. HS were applied successfully in many areas such as computer science, electrical engineering, civil Engineering, mechanical engineering and biomedical application as shown in Table III. So, based on all this consideration we will use HSA for feature selection in HCR as our future work.

TABLE III. SUMMARIZATION OF APPLICATION OF HARMONY SEARCH ALGORITHM

Area	Application of HSA	
	Authors. (Year)	***Tittle***
Computer Sciences	Jaco Fourie, Steven Mills, Richard Green. (2010)	Harmony filter: A robust visual tracking system using the improved harmony search algorithm
	Ebrahim Yazdi, Abolfazl Toroghi Haghighat. (2010)	Evolution of Biped Walking Using Neural Oscillators Controller and Harmony Search Algorithm Optimizer
	Erik Cuevas et. al (2010)	Circle Detection by Harmony Search Optimization
	M. Tamer Ayvaz (2010)	A linked simulation–optimization model for solving the unknown groundwater pollution source identification problems
Civil Engineering	Ali HaydarKayhan et al. (2011)	Selecting and scaling real ground motion records using harmony search algorithm
	Joong Hoon Kim, Zong Woo Geem, and Eung Seok Kim (2001)	Parameter estimation of the nonlinear muskingum model using harmony search
Electrical Engineering	A. Vasebi et al. (2007)	Combined heat and power economic dispatch by harmony search algorithm
	B. Majidi et al. (2008)	Harmonic Optimization in Multi-Level Inverters using Harmony Search Algorithm
	Rong Zhang and Lajos Hanzo (2009)	Iterative Multiuser Detection and Channel Decoding for DS-CDMA Using Harmony Search.
	Sukayapong Ngonkham and Panhathai Buasri (2009)	Harmony search algorithm to improve cost reduction in power generation system integrating large scale wind energy conversion system.
Mechanical Engineering	Parikshit Yadav et. al (2011)	An Improved Harmony Search algorithm for optimal scheduling of the diesel generators in oil rig platforms
Mechanical Engineering	Min Huang et. al (2009)	Guided Variable Neighborhood Harmony Search for Integrated Charge Planning in Primary Steelmaking Processes
	Zong Woo Geem and Han Hwangbo (2006)	Application of Harmony Search to Multi-Objective Optimization for Satellite Heat Pipe Design
	M. Fesanghary et. al (2009)	Design optimization of shell and tube heat exchangers using global sensitivity analysis and harmony search algorithm
	O. Zarei et. al (2009)	Optimization of multi-pass face-milling via harmony search algorithm
Bio & Medical Application	Abdulqader M. Mohsen et. al (2008)	HSRNAFold: A Harmony Search Algorithm for RNA Secondary Structure Prediction Based on Minimum Free Energy
	Osama Moh'd Alia et. al (2009)	Harmony Search-based Cluster Initialization for Fuzzy C-Means Segmentation of MR Images
	Jyotshna Dongardive et. al (2010)	Finding Motifs Using Harmony Search
	Sungho Muna and Zong Woo Geem (2010)	Determination of individual sound power levels of noise sources using a harmony search algorithm

VIII. CONCLUSION

This paper has presented a review about HCR in general and specifically concentrating on feature extraction and selection. The current trend on feature extraction and selection were discussed briefly. We also were investigating on metaheuristic algorithm which is harmony search algorithm as an optimization tool. Finally, a detailed and complete reference section has been provided. At the end of this, the result of this paper will be applied in feature extraction and selection in HCR using harmony search algorithm as our further study.

ACKNOWLEDGMENT

The authors honorably appreciate to Ministry of High Education Malaysia (MOHE), Universiti Teknologi Malaysia (UTM), Research Management Centre (RMC) with grant vote number 4F264, ZAMALAH UTM and Soft Computing Research Group (SCRG) for their support.

REFERENCES

[1] Shi-Fei Ding, Wei-KuanJia, Chun-Yang Su; Zhong-Zhi Shi, 2008. Research of pattern feature extraction and selection. Machine Learning and Cybernetics, International Conference on Volume 1, 12-15 July Page(s):466 – 471

[2] Dewi Nasien, Habibollah Haron, Siti Sophiayati Yuhaniz, "The Heuristic Extraction Algorithms for Freeman Chain Code of Handwritten Character", International Journal of Experimental Algorithms-IJEA, Vol. 1, Issue 1, pages 1-20.

[3] Plamondon, R., and Srihari, S. N. (2000). On-Line and Off-Line Handwriting Recognition: A Comprehensive Survey. IEEE Transactions on Pattern Analysis and Machine Intelligence, 22(1), 63-84.

[4] Arica, N., and Yarman-Vural. F.T., "An Overview of Character Recognition Focused on Off-line Handwriting", IEEE Trans. On Systems, Man, and Cybernetics, Vol. 31. No. 2, 2001, pp. 216-233.

[5] Impedovo, S. (1994). Fundamentals in Handwriting Recognition. In Computer and Systems Sciences. London:Springer-Verlag. International Conference on Image Processing. 11-14 September 2005. IEEE. II- 542-5.

[6] G.E.M.D.C. Bandara, S.D. Pathirana, R.M. Ranawana, "Use of fuzzy feature descriptions to recognize handwritten alphanumeric characters," in 1st Conference on Fuzzy Systems and Knowledge Discovery, Singapore, 2002.

[7] Paul D. Gader, James M. Keller, Raghu Krishnapuram, Jung-Hsien Chiang, Magdi A. Mohamed, "Neural and fuzzy methods in handwriting recognition," Computer, vol. 30, no. 2, pp. 79-86, Feb., 1997.

[8] S.Mori,C.YSuen,KYamamoto,HistoricalreviewofOCRresearchand development, ProceedingsoftheIEEE80(1992)1029–1058.

[9] C.Y Suen, C. Nadal, R Legault, T. A Mai, L Lam, Computer recognition of unconstrained handwrittennumerals,ProceedingsofIEEE80(1992) 1162–1180.

[10] Dewi Nasien, Siti S. Yuhaniz, Habibollah Haron, 2010. Recognition of Isolated Handwritten Latin Characters using One Continuous Route of Freeman Chain Code Representation and Feedforward Neural Network Classifier. International Science Index Vol:4, No:7, pp. 475-481.

[11] A. Suliman, M. N. Sulaiman, M. Othman and R. Wirza, 2010. Chain Coding and Pre Processing Stages of Handwritten Character Image File. electronic Journal of Computer Science and Information Technology (eJCSIT), Vol. 2, No. 1, pp. 6-13

[12] Serra, J., "Morphological Filtering : An Overview", SignalProcess, vol. 38, no. 1, 1994, pp. 3-11.

[13] Sonka, M., Hlavac V. and Boyle, R., Image Processing, Analysis and Machine Vision, 2nd ed. Pacific Grove CA:Brooks/Cole, 1999.

[14] Legault R. and Suen C.Y., "Optimal local weighted averaging methods in contour smoothing", IEEE Trans. Pattern Anal. Machine Intell., vol.18, pp. 690-706, July, 1997.

[15] Leu, J.G., "Edge Sharpening through ramp width reduction", Image Vis. Comput., vol. 18, no. 6-7, 2000, pp. 501-514.

[16] Solihin, Y. and Leedham C.G., , "Integral ratio: A new class of global thresholding techniques for handwriting images", IEEE Trans. Pattern Anal. Machine Intell., vol.21, August,1999, pp. 761-768.

[17] Lee, W.L. and Fan, K.C., "Document image preprocessing based on optimal Boolean filters", Signal Process, vol.80, no. 1, 2000, pp. 45- 55.

[18] Polesel, A., Ramponi, G., and Mathews, V., "Adaptive unsharp masking for contrast enhancements", in Proc. Int. Conf. Image Process., vol. 1, 1997, pp.267-271.

[19] Atici, A. and Yarman-Vural, F., "A Heuristic Method for Arabic Character Recognition", Signal Process, vol. 62, 2001, pp. 87-99.

[20] Chen, M.Y., Kundu, A. and Zhou, J., Off-Line Handwritten Word Recognition using HMM type Stochastic Network, IEEE Transactions of Pattern Analysis and Machine Intelligence, 16, 1994, 481-496.I. S. Jacobs and C. P. Bean, "Fine particles, thin films and exchange anisotropy," in Magnetism, vol. III, G. T. Rado and H. Suhl, Eds. New York: Academic, 1963, pp. 271–350.

[21] Reinhardt, J.M. and Higgins, W.E., 1996, Comparison between the morphological skeleton and morphological shape decomposition, IEEE Trans. Pattern Anal. Machine Intell., vol. 18, pp. 951-957, Sept.

[22] Yang, J. and Li, X.B., 1995, Boundary Detection using mathematical morphology, Pattern Recognition Letters, vol. 16, no.12, pp. 1287-1296.

[23] Kpalma, K., and Ronsin, J. (2007). An Overview of Advances of Pattern Recognition Systems in Computer Vision. In Vision Systems: Segmentation and Pattern Recognition (www.i-techonline.com). 169-194.

[24] Cole, L., Austin, D., and Cole, L. (2004). Visual Object Recognition Using Template Matching. Proceedings of the 2004 Australasian Conference on Robotics and Automation. Canberra. 6-8 December 2004.

[25] Roberto, B., and Tomaso, P. (1997). Template Matching: Matched Spatial Filters and Beyond. Pattern Recognition, 30(5), 751-768.

[26] Vapnik, V. N. (1999). An Overview of Statistical Learning Theory. Neural Networks, IEEE Transactions on, 10(5), 988-999.

[27] Jain, A. K., Duin, R. P. W., and Jianchang, M. (2000). Statistical Pattern Recognition: A Review. Pattern Analysis and Machine Intelligence, IEEE Transactions on, 22(1), 4-37.

[28] Duin, R. P. W., Roli, F., and de Ridder, D. (2002). A Note on Core Research Issues for Statistical Pattern Recognition. Pattern Recognition Letters, 23(4), 493- 499.

[29] Perlovsky, L. I. (1998). Conundrum of Combinatorial Complexity. Pattern Analysis and Machine Intelligence, IEEE Transactions on, 20(6), 666-670.

[30] Sarfraz, M. (2005). Computer-Aided Intelligent Recognition Techniques and Applications. England: John Wiley & Sons, Ltd

[31] I. S. Oh, J. S. Lee, C. Y. Suen, "Analysis of class separation and Combination of Class-Dependent Features for Handwriting Recognition", IEEE Trans. Pattern Analysis and Machine Intelligence, vol.21, no.10, pp.1089-1094, 1999.

[32] D. Trier, A. K. Jain, T. Taxt, "Feature Extraction Method for Character Recognition - A Survey", Pattern recognition, vol.29, no.4, pp.641-662, 1996.

[33] Hanson M, Powell H, Barth A, et al. Body area sensor networks: challenges and opportunities. IEEE Computer, 2009: 58 66

[34] Chipara, Lu C, Bailey T, et al. Reliability clinical monitoring using wireless sensor networks: experiences in a step-down hospital unit. SenSys 2010

[35] Fei Z. Research on Chinese pulse theory. Shanghai University of Traditional Chinese Medicine Press, 2005

[36] Lukman S, He Y, Hui S. Computational methods for traditional Chinese medicine: a survey. Computer Methods and Programs in Biomedicine, 2007:283-293

[37] LI Lei, ZHANG Li-liang, SU Jing-fei, 2012. Handwritten character recognition via direction string and nearest neighbor matching. The Journal of China Universities of Posts and Telecommunications. October 2012, 19(Suppl. 2): 160–165.

[38] S.V. Rajashekararadhya, P. Vanaja Ranjan, "A Novel Zone Based Feature Extraction Algorithm for Handwritten Numeral Recognition of Four Indian Scripts", Digital Technology, Journal, Vol. 2, pp. 41-51, 2009.

[39] Sandhya Arora et al. "Recognition of Non-Compound Handwritten Devnagari Characters using a Combination of MLP and Minimum Edit Distance", International Journal of Computer Science and Security (IJCSS), Volume-4, Issue 1.

[40] D. Trier, A. K. Jain, T. Taxt, "Feature Extraction Method for Character Recognition - A Survey", Pattern recognition, vol.29, no.4, pp.641-662, 1996.

[41] Santanu Chaudhury, Geetika Sethi, Anand Vyas, Gaurav Harit, "Devising Interactive Access Techniques for Indian Language Document Images", (ICDAR 2003).

[42] U. Pal, T. Wakabayashi, F. Kimura, "Comparative Study of Devnagari Handwritten Character Recognition using Different Feature and Classifiers", 10th Intl. Conf. on Document Analysis and Recognition, pp. 1111-1115, 2009.

[43] Santanu Chaudhury, Geetika Sethi, Anand Vyas, Gaurav Harit, "Devising Interactive Access Techniques for Indian Language Document Images", (ICDAR 2003).

[44] Tapan K Bhowmik, Swapan K Parui Utpal Roy, "Discriminative HMM Training with GA for Handwritten Word Recognition", IEEE, 2008.

[45] B.V.Dhandra, Mallikarjun Hangarge, "Global and Local Features Based Handwritten Text Words and Numerals Script Identification", Intl. Conf. on Computational Intelligence and Multimedia Applications, PP 471-475. 2007.

[46] Latesh Malik, P.S. Deshpande , "Recognition of printed Devnagari characters with regular expression in finite state models", International workshop on machine intelligence research, GHRCE Nagpur, India, 2009.

[47] M. Hanmandlu, O.V. Ramana Murthy, Vamsi Krishna Madasu, "Fuzzy Model based recognition of handwritten Hindi characters", Digital Image Computing Techniques and Applications 0-7695-3067-IEEE. Feb-07.

[48] Reena Bajaj, Lipika Dey , Santanu Chaudhury, "Devnagari numeral recognition by combining decision of multiple connectionist classifiers", Sadhana Vol. 27, Part 1, pp. 59–72, February 2002.

[49] Canasai Kruengkrai, Virach Sornlertlamvanich, Hitoshi Isahara, " Language, Script, and Encoding Identification with String Kernel Classifiers", Thai Computational Linguistics Laboratory, Thailand.

[50] P. S. Deshpande, Latesh Malik, Sandhya Arora," Recognition of Hand Written Devnagari Characters with Percentage Component Regular Expression Matching and Classification Tree", IEEE, 2007.

[51] Prachi Mukherji, Priti P. Rege, "Shape Feature and Fuzzy Logic Based Offline Devnagari Handwritten Optical Character Recognition", Journal of Pattern Recognition Research 4 (2009), pp 52-68.

[52] L. Heutte, J. V. Moreau, T. Paquet, Y. Lecourtier, and C. Olivier, " Combining structural and statistical features for the recognition of handwritten characters," Proceedings of 13th International Conference on Pattern Recognition, Vienna, Austria, 1996, Vol. 2, pp. 210-214.

[53] S. Arora, D. Bhattacharjee, M. Nasipuri, D. K. Basu and M. Kundu, "Combining multiple feature extraction techniques for handwritten Devnagari character recognition," IEEE Region 10 Colloquium and 3rd International Conference on Industrial and Information Systems, Dec. 2008.

[54] Y. Kimura, A. Suzuki, K. Odaka, "Feature selection for character recognition using genetic algorithm," IEEE Fourth International Conference on Innovative Computing, Information and Control (ICICIC), Kaohsiung , pp. 401-404, Dec. 2009.

[55] Vamvakas, G., B. Gatos, and S.J. Perantonis, Handwritten character recognition through two-stage foreground sub-sampling. Pattern Recognition, 2010. 43(8): p. 2807-2816.

[56] Chacko, B., et al., Handwritten character recognition using wavelet energy and extreme learning machine. International Journal of Machine Learning and Cybernetics, 2012. 3(2): p. 149-161.

[57] Wang, X. and A. Sajjanhar, Polar Transformation System for Offline Handwritten Character Recognition, in Software Engineering, Artificial Intelligence, Networking and Parallel/Distributed Computing 2011, R. Lee, Editor. 2011, Springer Berlin Heidelberg. p. 15-24.

[58] Yang, Y., X. Lijia, and C. Chen, English Character Recognition Based on Feature Combination. Procedia Engineering, 2011. 24(0): p. 159-164.

[59] Chel, H., A. Majumder, and D. Nandi, Scaled Conjugate Gradient Algorithm in Neural Network Based Approach for Handwritten Text Recognition, in Trends in Computer Science, Engineering and Information Technology, D. Nagamalai, E. Renault, and M. Dhanuskodi, Editors. 2011, Springer Berlin Heidelberg. p. 196-210.

[60] AlKhateeb, J.H., et al., Offline handwritten Arabic cursive text recognition using Hidden Markov Models and re-ranking. Pattern Recognition Letters, 2011. 32(8): p. 1081-1088.

[61] Rajput, G. and R. Horakeri, Handwritten Kannada Vowel Character Recognition Using Crack Codes and Fourier Descriptors, in Multi-disciplinary Trends in Artificial Intelligence, C. Sombattheera, et al., Editors. 2011, Springer Berlin Heidelberg. p. 169-180.

[62] Sharma, D. and P. Jhajj, Comparison of Feature Extraction Methods for Recognition of Isolated Handwritten Characters in Gurmukhi Script, in Information Systems for Indian Languages, C. Singh, et al., Editors. 2011, Springer Berlin Heidelberg. p. 110-116.

[63] Choudhary, A., R. Rishi, and S. Ahlawat, Unconstrained Handwritten Digit OCR Using Projection Profile and Neural Network Approach, in Proceedings of the International Conference on Information Systems Design and Intelligent Applications 2012 (INDIA 2012) held in Visakhapatnam, India, January 2012, S. Satapathy, P.S. Avadhani, and A. Abraham, Editors. 2012, Springer Berlin Heidelberg. p. 119-126.

[64] Li, L., L.-l. Zhang, and J.-f. Su, Handwritten character recognition via direction string and nearest neighbor matching. The Journal of China Universities of Posts and Telecommunications, 2012. 19, Supplement 2(0): p. 160-196.

[65] Nemouchi, S., L. Meslati, and N. Farah, Classifiers Combination for Arabic Words Recognition: Application to Handwritten Algerian City Names, in Image and Signal Processing, A. Elmoataz, et al., Editors. 2012, Springer Berlin Heidelberg. p. 562-570.

[66] Ahmed, I., S. Mahmoud, and M. Parvez, Printed Arabic Text Recognition, in Guide to OCR for Arabic Scripts, V. Märgner and H. El Abed, Editors. 2012, Springer London. p. 147-168.

[67] Likforman-Sulem, L., et al., Features for HMM-Based Arabic Handwritten Word Recognition Systems, in Guide to OCR for Arabic Scripts, V. Märgner and H. El Abed, Editors. 2012, Springer London. p. 123-143.

[68] Kessentini, Y., T. Paquet, and A. Ben Hamadou, Multi-stream Markov Models for Arabic Handwriting Recognition, in Guide to OCR for Arabic Scripts, V. Märgner and H. El Abed, Editors. 2012, Springer London. p. 335-350.

[69] Bhattacharya, U., et al., Offline recognition of handwritten Bangla characters: an efficient two-stage approach. Pattern Analysis and Applications, 2012. 15(4): p. 445-458.

[70] Reddy, G.S., et al. Combined online and offline assamese handwritten numeral recognizer. in Communications (NCC), 2012 National Conference on. 2012.

[71] Muhammad Naeem Ayyaz, Imran Javed and Waqar Mahmood. Handwritten Character Recognition Using Multiclass SVM Classification with Hybrid Feature Extraction. Pak. J. Engg. & Appl. Sci. Vol. 10, Jan., 2012 (p. 57-67)

[72] Vidya V, Indhu T R, Bhadran V K,R Ravindra Kumar, "Malayalam Offline Handwritten Recognition using Probabilistic Simplified Fuzzy ARTMAP", Advances in Intelligent Systems and Computing Volume 182, 2013, pp 273-283.

[73] Primekumar K.P,Sumam Mary Idiculla, "On-line Malayalam Handwritten Character Recognition using HMM and SVM", 2013 International Conference on Signal Processing, Image Processing and PatternRecognition[ICSIPR].

[74] C. De Stefano, F. Fontanella , C. Marrocco, A. Scotto di Freca, 2014. A GA-based feature selection approach with an application to handwritten character recognition. Pattern Recognition Letters 35 (2014) 130–141.

[75] I. Guyon and A. Elisseeff. An introduction to variable and feature selection. J. Mach. Learn. Res., 3:1157 1182, 2003.

[76] Kwak, N., Choi, C.-H., 2002. Input feature selection for classification problems. IEEE Trans. Neural Networks 13 (1), 143–159.

[77] Somol, P., Pudil, P., Kittler, J., 2004. Fast branch and bound algorithms for optimal feature selection. IEEE Trans. Pattern Anal. Machine Intell. 26 (7), 900–912.

[78] Goldberg, D.E., 1989. Genetic Algorithms in Search Optimization and Machine Learning. Addison-Wesley.

[79] M. Kudo and J. Sklansky. Comparison of algorithms that select features for pattern recognition. Pattern Recognition, 33(1):25–41, 2000.

[80] J.-S. Lee, I.-S. Oh, and B.-R. Moon. Hybrid genetic algorithms for feature selection. IEEE Trans. Pattern Anal. Mach. Intell., 26(11):1424–1437, 2004.

[81] J. Yang and V. Honavar. Feature subset selection using a genetic algorithm. IEEE Intelligent Systems, 13:44–49,1998.

[82] Kudo, M., Sklansky, J., 2000. Comparison of algorithms that select features for pattern recognition. Pattern Recognition 33 (1), 25–41.

[83] Oh, I.-S., Lee, J.-S., Moon, B.-R., 2004. Hybrid genetic algorithms for feature selection. IEEE Trans. Pattern Anal. Machine Intell. 26 (11), 1424–1437.

[84] Cordella, L., De Stefano, C., Fontanella, F., Marrocco, C., Scotto di Freca, A., 2010. Combining single class features for improving performance of a two stage classifier. In: 20th Internat. Conf. on Pattern Recognition (ICPR 2010), pp. 4352– 4355.

[85] De Stefano, C., Fontanella, F., Marrocco, C., Schirinzi, G., 2007. A feature selection algorithm for class discrimination improvement. In: Geoscience and Remote Sensing, Symposium, 2007 (IGARSS07), pp. 425 428.

[86] Siedlecki, W., Sklansky, J., 1989. A note on genetic algorithms for large-scale feature selection. Pattern Recognition Lett. 10 (5), 335–347.

[87] Yang, J., Honavar, V., 1998. Feature subset selection using a genetic algorithm. IEEE Intell. Systems 13, 44 49.

[88] Dewi Nasien, Habibollah Haron and Siti S. Yuhaniz. (2010). Metaheuristics Methods (GA & ACO) For Minimizing the Length of Freeman Chain Code from Handwritten Isolated Characters, World Academy of Science Engineering and Technology, Vol. 62, February 2010, ISSN: 2070-3274, Article 41, pp. 230-235

[89] Reza Azmi , Boshra Pishgoo , Narges Norozi , Maryam koohzadi , Fahimeh baesi,2010. A hybrid GA and SA algorithms for f eature selection in recognition of hand-printed Farsi characters.

[90] A. Marcano-Cedeño, J. Quintanilla-Domínguez, M.G. Cortina-Januchs and D. Andina,2010. Feature Selection Using Sequential Forward Selection and classification applying Artificial Metaplasticity Neural Network.

[91] Dewi Nasien, Habibollah Haron, Siti Sophiayati Yuhaniz. (2011). The Heuristic Extraction Algorithm for Freeman Chain Code of Handwritten Character. International Journal of Experimental Algorithms (IJEA). Publisher: CSC Press, Computer Science Journals, Volume: 1, Issue: 1, pp. 1-20, ISSN: 2180-1282.

[92] Gheith Abandah and Nasser Anssari. 2009. Novel Moment Features Extraction for Recognizing Handwritten Arabic Letters Journal of Computer Science 5 (3): 226-232, 2009 ISSN 1549-3636. Science Publications

[93] Nibaran Das, Ram Sarkar, Subhadip Basu, Mahantapas Kundu, Mita Nasipuri, Dipak Kumar Basu.2012. A genetic algorithm based region sampling for selection of local features in handwritten digit recognition application. Applied Soft Computing 12 (2012) 1592–1606.

[94] Abhinaba Roy, Nibaran Das, Ram Sarkar, Subhadip Basu, Mahantapas Kundu, Mita Nasipuri,2012. Region Selection in Handwritten Character Recognition using Artificial Bee Colony Optimization. 2012 Third International Conference on Emerging Applications of Information Technology (EAIT). pp. 183-186.

[95] Nagasundara K B, Guru D S, Manjunath S, 2012. Feature selection and Indexing of Online Signatures.

[96] C. De Stefano, F. Fontanella, C. Marrocco, A. Scotto di Freca,2014. A GA-based feature selection approach with an application to handwritten character recognition. Pattern Recognition Letters 35 (2014) 130-141.

[97] Abhinaba Roy, Nibaran Das, Ram Sarkar, Subhadip Basu, Mahantapas Kundu, Mita Nasipuri,2014. An Axiomatic Fuzzy Set Theory Based Feature Selection Methodology for Handwritten Numeral Recognition. ICT and Critical Infrastructure: Proceedings of the 48th Annual Convention of Computer Society of India- Vol I Advances in Intelligent Systems and Computing Volume 248, 2014, pp 133-140.

[98] F. Ghareh Mohammadi, M.Saniee Abadeh, 2014. Imagesteg analysis using a bee colony based feature selection algorithm. Engineering Applications of Artificial Intelligence 31(2014)35-43.

[99] Tou, J. T., and Gonzalez, R. C. (1972). Recognition of Handwritten Characters by Topological Feature Extraction and Multilevel Categorization. Computers, IEEE Transactions on, C-21(7), 776-785.

[100] Sohel, F. A., Karmakar, G. C., and Dooley, L. S. (2007). Bezier Curve-Based Character Descriptor Considering Shape Information. Proceedings of the 2007 IEEE International Conference on Computer and Information Science. 11-13 July 2007. IEEE. 212-216.

[101] Jähne, Bernd; Digital Image Processing: Concepts, Algorithms, and Scientific Applications. Edition: 6, Published by Springer, 2005.

[102] M. Sonka, V. Hlavac and R. Boyle, Image Processing Analysis and Machine Vision, Second Edition, PWS, (1999).

[103] P. Meer, C.A. Sher and A. Rosenfeld, The chain pyramid: Hierarchical contour processing, IEEE Pattern Analysis and Machine Intelligence 12, 363-375, (1990).

[104] Kormos. J., Vereb. K. Recognition of Chain-Coded Patches With Statistical Methods. Mathematical and Computer Modelling 38 (2003) 903-907.

[105] Liu, Y.K., Zalik, B.: An efficient chain code with Huffman coding, Pattern Recognition, 38(4), 2005, 553-557.

[106] Sánchez-Cruz, Hermilo., Bribiesca, Ernesto., Rodríguez- Dagnino, R.M. Efficiency of Chain Codes to Represent Binary Objects. Volume 40, Issue 6, June 2007, Pages 1660-1674

[107] Wulandhari. L.A., Haron Habibolah. The Evolution and Trend of Chain Code Scheme. ICGST-GVIP, ISSN 1687- 398X, Volume (8), Issue (III), October 2008.

[108] K.Palagyi, Shape representation/description, In Proceedings of the 8th SSIP, Zagreb, Croatia, (2000).

[109] Ren Mingwu, Yang Jingyu and Sun Han. Tracing Boundary Contours in Binary Image, Image and Vision Computing. Volume 20, Issue 2, 2002, Pages 125-13.

[110] Cabrelli C A, Molter U M, Automatic Representation of Binary Images, IEEE Transactionon Pattern and machine Intelligence 12, 1990, 1190-1196.

[111] Freeman. H, Techniques for the Digital Computer Analysis of Chain-Encoded Arbitrary Plane Curves, Proc. Natn. Electron. Conf. 18 (1961) 312-324.

[112] Freeman H, Computer Processing of Line-Drawing Images, ACM Computing Surveys 6, 1974, 57-97.

[113] Haron, Habibollah; Dzulkifli, Mohammad; Shamsuddin, S.S. Enhancement Algorithms for 3D Object Interpreter. PSc Thesis, Universiti Teknologi Malaysia, 2004.

[114] Fadoul, F.M; Development of Mapping and Visualizing Algorithm of Vertex Chain Code from Thinned Binary Image, MSc Thesis, Universiti Teknologi Malaysia, 2008.

[115] Paul Kwok. A thinning algorithm by contour generation. Source Communications of the ACM archive Volume 31, Issue 11 (November 1988). Pages: 1314 – 1324.

[116] Liu, Y. K. and Zalik, B. "An efficient chain code with huffman coding". Pattern Recognition, 38(4):553-557, 2005

[117] Sánchez-Cruz, H. Bribiesca, E. and Rodríguez-Dagnino, R. M. "Efficiency of chain code to represent binary objects". Pattern Recognition, 40(6):1660-1674, 2007

[118] Wulandhari, L. A. and Haron, H. "The evolution and trend of chain code scheme". Graphics, Vision and Image Processing, 8(3):17-23, 2008

[119] Gebrail Bekdas. Harmony Search Algorithm Approach for Optimum Design of Post-Tensioned Axially Symmetric Cylindrical Reinforced ConcreteWalls. J Optim Theory Appl, DOI 10.1007/s10957-014-0562 2, 2013

[120] Holland, J.H.: Adaptation in Natural and Artificial Systems. University of Michigan Press, Ann Arbor (1975)

[121] Impedovo, S., Ottaviano, L. and Occhinegro, S., 1991, Optical Character Recognition – A Survey, International Journal of Pattern Recognition & Artificial Intelligence, 5, 1-24.

[122] Zhaoqi, B. and Xuegong, Z. "Pattern Recognition", 2nd Edition, Tsinghua University Press., (2000).

Resource Provisioning in Single Tier and Multi-Tier Cloud Computing: "State-of-the-Art"

Marwah Hashim Eawna
Faculty of Computer and Information Sciences
Ain Shams University
Cairo, Egypt

Salma Hamdy Mohammed
Faculty of Computer and Information Sciences
Ain Shams University
Cairo, Egypt

El-Sayed M. El-Horbaty
Faculty of Computer and Information Sciences
Ain Shams University
Cairo, Egypt

***Abstract*—Cloud computing is a new computation trend for delivering information as long as an electronic device needs to access of a web server. One of the major pitfalls in cloud computing is related to optimizing the resource provisioning and allocation. Because of the uniqueness of the model, resource provisioning is performed with the objective of minimizing time and the costs associated with it. This paper reviews the state-of-the-art of managing resources of the cloud environments in theoretical research. This study discusses the performance and analysis for well-known cloud provisioning resources techniques, single tier and multi-tier.**

***Keywords*—Cloud Computing; Resource Provisioning**

I. INTRODUCTION

The cloud environment describes a company, organization or individual that uses a web-based application for every task rather than installing software or storing data on a computer. All cloud environments are not common but a move toward this is a long-term goal for cloud computing enthusiasts and cloud capitalists.

Many challenges are influenced to adopt the cloud computing technology such as security, resources allocation, resources provisioning and others. In provisioning resource for cloud computing environment the major challenge is to determine the right amount of resources required for the execution of work in order to minimize the financial cost from the perspective of users and to maximize the resource utilization from the perspective of service providers [1].

The resource provisioning must meet Quality of Service (QoS) parameters like availability, throughput, response time, security, reliability etc., and thereby avoiding Service Level Agreement (SLA) violation [2]. There are entirely two generic way of resource provisioning, Static and dynamic.

1) Static Resource Provisioning: usually provides the peak time needed resource all the time for the application. In this kind of provisioning most of the time the waste of resource due to workload is not in a peak, but resource providers provide the maximum required resource to prevent SLA violation.

2) Dynamic Resource Provisioning: the basic fundamental idea in the latter way is providing the resources based on the application needs, this helps the provider to assign the Non-loaded resources (which become free to used now) to the new users. This method reduces a fraction of providers' development costs by utilizing current available resources and beside that the user can happily just pay for the amount of the resources which were really used [3]. Moreover, the parameters of resource provisioning is presented as:

1) Response time: The resource provisioning algorithm designed must take minimal time to respond when executing the task.

2) Minimize Cost: From the Cloud user point of view cost should be minimized.

3) Revenue Maximization: This is to be achieved from the Cloud Service Provider's view.

4) Fault tolerant: The algorithm should continue to provide service in spite of failure of nodes.

5) Reduced SLA Violation: The algorithm designed must be able to reduce SLA violation.

6) Reduced Power Consumption: virtual machine placement & migration techniques must lower power consumption [4].

The rest of this paper is organized as follows: Sections 2 contains a review of resource provisioning in single tier and multi-tier cloud environments. Section 3 illustrates the most popular efficiency of Single tier and Multi-tier architectures. Finally, Conclusions and future works are given in Section 4.

II. RELATED WORK

This section explains two basic architectures are dependent in the resources provision in cloud computing environment like single tier and multi-tier architecture.

A. Single-tier technique

A single-tier architecture of cloud computing has a set of servers that are used to provide resources by receiving

requests from user in presentation server and looking information in application server and store the information in database server.

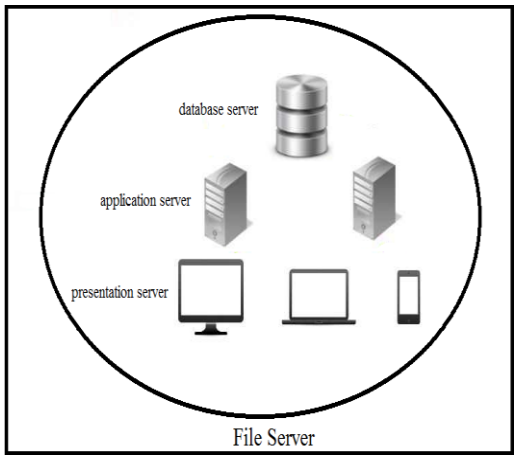


Fig. 1. Single tier architecture in cloud computing

As show in figure1, single tier architecture there are no interactive between the servers and is not allowed to share resources between server that cause consumption of resources and loss of time when compared with multi-tier architecture.

Some of the existing techniques involving nature-inspired meta-heuristics have become the new focus of resource allocation. For example, Tarun goyal et al. [5] Scheduling a model based on minimum network delay using Suffrage Heuristic coupled with Genetic algorithms for scheduling sets of independent jobs algorithm is proposed, the result show that the schedule of multiple jobs on multiple machines in an efficient manner such that the jobs take the minimum time for completion.

Othman et al. [6] proposed a novel Simulated Annealing (SA) algorithm for scheduling tasks in cloud environments. SA exploits an analogy between the way in which a metal cools and freezes into a minimum energy crystalline structure (the annealing process) and the search for a minimum in a more general system. Results show that this approach for job scheduling not only guarantees the QoS requirement of customer job but also ensures to make best profit of cloud providers. It has also concern about real execution time of jobs in different systems as well as deadline and penalty cost in the algorithm.

Shaobin Zhan et al. [7] introduces an improved Particle Swarm Optimization (IPSO) algorithm in resources scheduling strategy by Add simulated annealing into every iteration of PSO, through experiments, the results show that this method can reduce the task average running time, and raises the rate availability of resources.

Talwinder Kaur et al. [8] Improved Particle Swarm Optimization (IPSO), Simulated Annealing (SA) Algorithm, and Hybrid Particle Swarm Optimization-Simulated Annealing algorithms based on utilizing and scheduling resources. The experiment show that by using this algorithm provisioning resource in very less time as compared to the existing algorithm.

Xiaotang Wen el al. [9] improved algorithm to provide resource by combine Ant Colony Optimization (ACO) with particle swarm optimization algorithm to improve the efficiency of resource scheduling in cloud computing environments. Table 1 summarizes the advantages and disadvantages of single tier mechanisms are surveyed in this section.

TABLE I. RESOURCE PROVISIONING ALGORITHMS IN SINGLE TIER TECHNIQUES

Parameters	Techniques	Attributes	Authors
Time	Scheduling model based on GA	minimize the make span	Tarun goyal et al. [5]
	SA-based approach for sche-duling in cloud envi-ronment	minimize execution time	Monir Abdullah et al. [6]
	PSO algorithm in resources scheduling strategy	reduce the task average running time, and raises the rate availability of resources	Shaobin Zhan et al. [7]
	Use PSO and SA algorithm	less execution time as compare with existing algorithm	Talwinder et al.[8]
	Use PSO and ACO algorithm	Efficiently and speed to provide resource	Xiaotang et al.[9]

B. Multi-tier technique

This architecture partitions the application process into multiple tiers. Each tier provides certain functionality. The benefit of such architecture is that it can provide a high level of scalability and reliability. However, the resource allocation among these tiers will be more difficult due to the interdependency between the tiers. A multi-tier cloud computing application may span multiple nodes. Specifically, most multi-tier cloud computing applications use 3-tier architecture.

As shown in figure 2, the first tier, named presentation tier, consists of Web servers. It displays what is presented to the user on the client side within their Web browsers. For the Web server tier, it mainly has three functions:

1) Admitting/denying requests from clients and services static Web requests.

2) Passing requests to the Application server.

3) Receiving response from Application server and sending them back to clients. Examples of Web servers include Apache Server and Microsoft Internet Information Server (IIS).

The second tier, named business tier, consists of Application servers. Business logic processing is performed at this tier. There are also three functions at the Application server tier:

1) Receiving requests from the Web server.

2) Looking up information in the database and processing the information.

3) Passing the processed information back to the Web server.

The last tier, named data tier, consists of database servers. It handles database processing and data accessing. Database server tier is used to store and retrieve a Web site's information (e.g., user accounts, catalogs to reports, and customer orders) [10].

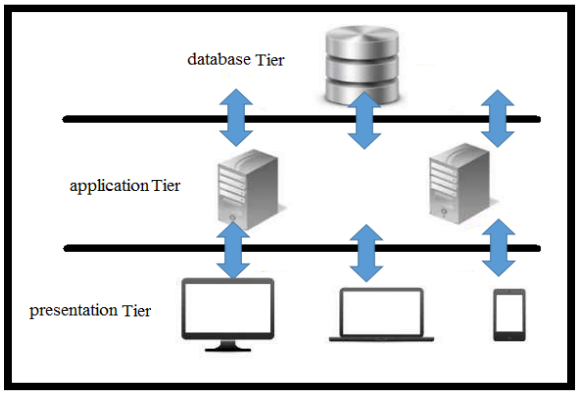


Fig. 2. Multi-tier architecture of cloud computing

Hadi Goudarzi el al. [11] considered the problem of the resource allocation to optimize the total profit gained from the SLA contracts and lost from operational cost. They assume that servers are characterized by their maximum capacity in three dimensions: processing power, memory usage, and communication bandwidth. While guaranteeing SLAs for clients with applications that require multiple tiers of service to complete.

Bhuvan Urgaonkar et al. [12] propose a novel data center architecture based on virtual machine monitors to reduce provisioning overheads, this technique reduced the overhead of switching servers across applications from several minutes to less than a second, while meeting the performance targets of residual sessions.

Chandra at el. [13] considered the problem of resource allocation on shared data centers, where they modeled a server resource as a generalized process sharing server and used a time-domain description of the server to model transient system states. These techniques can judiciously allocate system resources, especially under transient overload conditions

Heng et al. [14] use a benefit-aware approach with feedback control theory to solve the problem of continuously guarantees the SLA in the new configuration in multi-tier application. This approach can reduce resource provisioning cost by as much as 30% compared with a cost oblivious approach, and can effectively reduce SLA violations compared with a cost-aware approach.

TABLE II. RESOURCE PROVISIONING ALGORITHMS IN MULTI-TIER TECHNIQUES

Parameters of	Techniques	Attributes	References
Time	novel dynamic provisioning technique	double the application capacity reduced the overhead of switching servers from several minutes to less than a second	Bhuvan Urgaonkar et al. [11]
SLA	model for SLA-based multi-dimensional resource allocation scheme	meet SLA and effectiveness	Hadi Goudarzi el al. [12]
	novel benefit-aware provisioning approach	effective in reducing both cost and SLA violations	Heng et al. [13]
	model the server resource by use a time-domain description of (GPS) server	judiciously allocate system resources	Abhishek .[14]

Table 2 summarizes resource provisioning in multi-tier technique, these techniques considered SLA and real execution time of job in different system as well as soft deadline and penalty cost in the algorithm. Table2 ensures that prevent SLA violated and give a good profit for the different cloud provider.

III. EFFICIENCY OF SINGLE TIER AND MULTI-TIER ALGORITHMS

There are several algorithm has been developed to provide a better scheduling in a single tier and multi-tier cloud environment. Experiment by using CloudSim shows that the improved algorithm not only accelerated the convergence speed, but also avoided falling into local optimum solution, and achieved the purpose that the user tasks were efficiently provided appropriate resources in cloud computing, which improved the resource utilization ratio.

As shown in figure 3, the algorithm that combine between PSO and ACO takes time more than other algorithm that combine between PSO and SA.

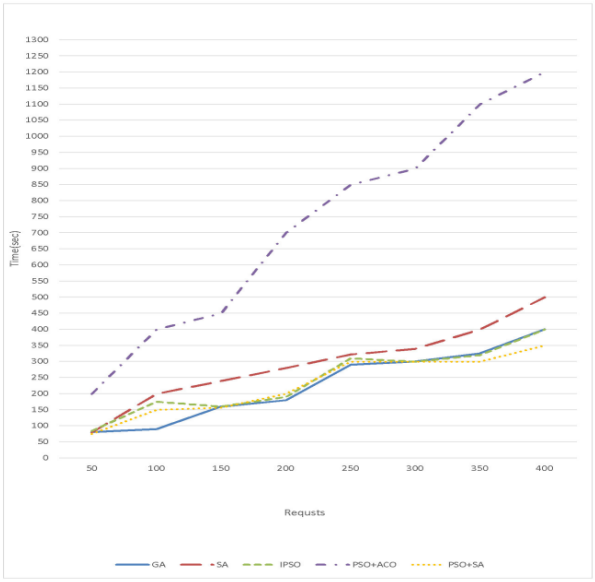


Fig. 3. Comparison among algorithms of single tier architecture

For instance, if the number of requests is 400 in the algorithm that combine between PSO and ACO the complete execution time will be 1200 sec. and in the algorithm that combine between PSO and SA the complete execution time will be 350 sec.

Furthermore, the Pseudo code of the algorithm combine between PSO and SA is iullstarted in figure 4.

```
1: procedure PSO-SA algorithm
2: CalculateExecTimes();
3: initSwarm ();
4: initGlobalBest();
5: for i=0 to numberIterations do
6: for j=0 to numberParticles do
7: calculateInertiaValue();
8: calculateNewVelocities();
9: calculateNewPositions();
10: calculateFitnessValue();
11: evaluateSolution();
12: updateParticleMemory();
13: updateGlobalBest();
14.UpdateGlobalBestDependonSAalgorithm
15: end for
16: end for
17: end procedure
```

Fig. 4. Pseudo code of algorithm that combine between PSO and SA based resource provisioning

On the other side, the multi-tire algorithms are surveyed in [11], [12], [13], and [14] have been implemented based on CloudSim environment. Figure 5 shows the efficiency of these algorithms.

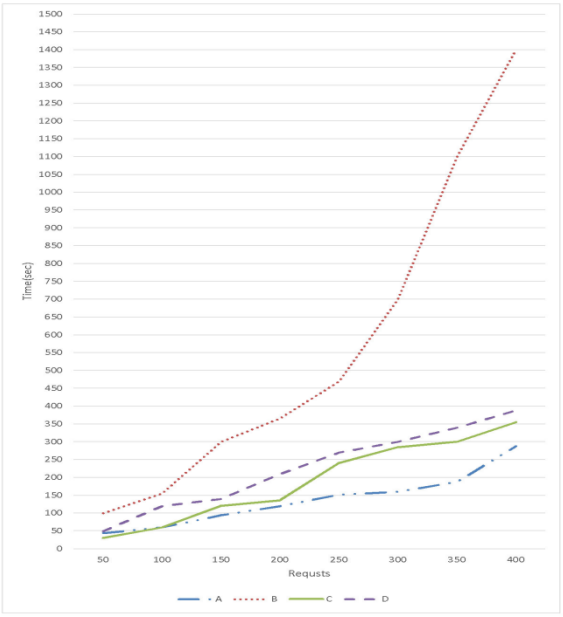


Fig. 5. Comparison among algorithms of multi-tier architecture

According to figure 5, the algorithm of novel dynamic provisioning technique [B] takes time more than others. However, less time is needed for the algorithms that use multi-dimensional SLA-based resource allocation [A]. Finally, figure 6 shows the Pseudo code of the multi-dimensional SLA algorithm in multi-tier technique**.**

Algorithm *Resource_Consolidate* ()
// Search the solution space to find better profit
TP = total profit;
Initialize the forces between clients and servers;
// calculate force differentials
$D_{i,j\to k}^{\alpha,t} = F_{ik}^{\alpha,t} - F_{ij}^{\alpha,t}$; $\forall j,i,t,\alpha$
ΔF = 1;
While (ΔF > 0) {
ΔF = max ($D_{ij\to k}^{\alpha,t}$); // client i and α
j = selected source server;
k = selected destination server type;
g = selected destination server;
If (ΔF is toward an ON server in server type k){
g = find the least busy server in k, assigned to tier t;
If (lower bound constraints satisfied) goto Re-Assign;
Else goto skip Re-Assign;}
Else If (ΔF is toward an OFF server in server type k){
g = find an OFF server in k;
If (found an OFF server) goto Re-Assign;
Else goto skip Re-Assign;}
Else If (ΔF is toward a server serving client i) goto Re-Assign;
Re-Assign: Re-assign α portion of the requests to g from j;
Update force related to j, g and client i;
P = total profit;
If (P>TP) TP = P; save the state;}
Skip Re-Assign: Update the move limit; }

Fig. 6. Pseudo code of the algorithm using multi-dimensional SLA-based resource provisioning

IV. Conclusions and Future Work

The investigations which were studied above are trying to optimize and utilize the resources. Several methods were mentioned here which used different parameters as a goal for resource provisioning such as response time, rejection rate, service level agreement (SAL) violation rate, cost etc. For provisioning planning should take appropriate provisioning times, Provisioning resources too soon will wastes our resources and therefore our money, on the other side provisioning resources too late will cause potentially SLA violations and makes the users angry.

Several ideas were reviewed and it can be concluded from them that managing such big resources for human administrators is not possible anymore and administrators are going to be replaced with managing systems. These systems must use techniques that are able to estimate and allocated resource in the most efficient way while avoiding SLA violations. New trends are needed with less human intervention so some new search techniques involving nature-inspired meta-heuristics have become the new focus of resource allocation research by using it provide very good solutions in a reasonable time. Provisioning resources by using multi-tier architecture that given continuously guarantees the SLA and provide a high level of scalability and reliability but the resource provisioning by multi-tier architecture is more difficult due to the fact that the resource demand at each tier is different. However, Single-tier architecture has relatively simple structure and is easy to setup.

So far, there are no attempts to use algorithms of meta-heuristic technique to provide resources in multi-tier architecture, our Improvement methodology will deals with the resource provisioning in multi-tier architecture in cloud computing by using of meta-heuristic technique such as SA, PSO, PSO-SA algorithm. Finally, compare our allocation techniques with other allocation techniques to evaluate their relative effectiveness.

Acknowledgements

This work was supported by Faculty of Computer and Information Sciences, Ain Shams University

References

[1] Eun-Kyu Byuna, Yang-Suk Keeb, Jin-Soo Kimc,Seungryoul Maenga, "Cost optimized provisioning of elastic resources for application workflows," *Future Generation Computer Systems,vol. 27,* pp. 1011–1026, 2011.

[2] Guruprasad, Bhavani B H and H S, "Resource Provisioning Techniques in Cloud Computing Environment: A Survey," International Journal of Research in Computer and Communication Technology, Vol 3, no. Issue3, pp. 395-400, March-2014.

[3] S. J. Hamid Reza Qavami, "A Survey On Resource Provisioning In Cloud Computing," International Journal of Research in Computer and Communication Technology, Vol.2, no. Issue.2, pp. 160-167, February 2014.

[4] Guruprasad, Bhavani B H and H S, "Resource Provisioning Techniques in Cloud Computing Environment: A Survey," . Vol 3, no. Issue3, pp. 395-401, March-2014.

[5] Agrawal, Tarun goyal & Aakanksha, "Host Scheduling Algorithm Using Genetic Algorithm In Cloud Computing Environment," international journal of research in engineering & technology (ijret), Vol. 1, no. Issue 1, pp. 7-12, june 2013.

[6] Othman, Monir Abdullah and Mohamed, "Simulated Annealing Approach To Cost-Based Multi- Quality Of Service Job Scheduling In Cloud Computing Enviroment," American Journal of Applied Sciences,vol.11, pp. 872-877, 2014.

[7] Huo, Shaobin Zhan and Hongying, "Improved PSO-based Task Scheduling Algorithm in Cloud Computing," Journal of Information & Computational Science, vol.9, pp. 3821–3829, 2012.

[8] Talwinder Kaur, Seema Pahwa. s.l., "An Upgraded Algorithm of Resource Scheduling using PSO and SA in Cloud Computing.," International Journal of Computer Applications, vol. 74, pp. 28-32, July 2013.

[9] Xiaotang Wen, Minghe Huang, Jianhua Shi. s.l., "Study on Resources Scheduling Based on ACO Algorithm and PSO Algorithm in Cloud Computing.," IEEE, pp. 219-222, 2012.

[10] Dong Huang, Bingsheng He and Chunyan Miao, "A Survey of Resource Management in Multi-Tier Web Applications," IEEE, vol. 16, no. 3, pp. 1574 - 1590, 29 January 2014.

[11] Pedram., Hadi Goudarzi and Massoud, "Multi-dimensional SLA-based Resource Allocation for Multi-tier Cloud Computing Systems," Cloud Computing (CLOUD), IEEE International Conference on, pp. 324 - 331. 2011.

[12] Bhuvan Urgaonkar, Prashant Shenoy, Abhishek Chandray, and Pawan Goyal, "Agile, Dynamic Provisioning of Multitier".

[13] Chandra, W. Gong, and P. Shenoy, "Dynamic Resource Allocation for Shared Data Centers Using Online Measurements," Proceedings of the 11th International Conference on Quality of Service, vol. 2707, pp. 381-398, 2003.

[14] Heng WU ,Wenbo ZHANG, Jianhua ZHANG, JunWEI, Tao HUANG, "A benefit-aware on-demand provisioning approach for multi-tier application in cloud computing," Frontiers of Computer Science, pp. 459–474, 2013.

Improving Web Movie Recommender System Based on Emotions

Karzan Wakil
Computer Department
Institute of Training and Educational Development-Sulaimani
Sulaimaniyah-Iraq

Rebwar Bakhtyar
Information Technology Department
Computer Science Institute-Sulaimani Polytechnic University
Sulaimaniyah-Iraq

Karwan Ali
Computer Science Department
University of Human Develpment
Sulaimaniyah-Iraq

Kozhin Alaadin
Computer Science Department
University of Human Develpment
Sulaimaniyah-Iraq

***Abstract*—Recommender Systems (RSs) are garnering a significant importance with the advent of e-commerce and e-business on the web. This paper focused on the Movie Recommender System (MRS) based on human emotions. The problem is the MRS need to capture exactly the customer's profile and features of movies, therefore movie is a complex domain and emotions is a human interaction domain, so difficult to combining together in the new Recommender System (RS). In this paper, we prepare a new hybrid approach for improving MRS, it consists of Content Based Filtering (CBF), Collaborative Filtering (CF), emotions detection algorithm and our algorithm, that presented by matrix. The result of our system provides much better recommendations to users because it enables the users to understand the relation between their emotional states and the recommended movies.**

Keywords—movie recommender system; collaborative filtering; content based filtering; emotion; CF; CBF; MRS

I. INTRODUCTION

RSs can be described as the software tools and techniques offering recommendations for items to be of use to a user [1]. Users today in the world with the internet and its associated information explosion are facing with the problematic situations that have too many options. Right from looking for a restaurant to looking for a good investment selection, there is huge information available. As assisting to the users as they can deal with this information burst, companies have organized RSs to direct them. The research in the area of RSs has been going on for several decades now, but the interest remaining elevation due to the plenty of practical applications and the problem rich domain. There are many active examples of online RSs that are implemented such as a RS for books at (Amazon.com), Library for movies at (MovieLens.org), CDs at (CDNow.com), and so on [2].

Recommender systems are now common both commercially and in the research community, where many methods and techniques have been suggested for providing recommendations. In many cases, a system designer that wishes to employ a RS must choose between a set of candidate approaches. A first step towards selecting the appropriate algorithm is to decide which properties of the application to focus upon when making this choice. Indeed, RSs have a variety of properties and attributes that may affect user experience, such as accuracy, robustness, scalability, and so forth [3]. Movie recommendation system which is recommends movies to users based on their personal information and their answers to questions based on movies [2].

The proposed system for the movie recommendation uses both CF and CBF. In order to provide reliable recommendation, the RSs need to capture the customer needs and preferences exactly into the user profile. However, for subjective and complex products such as movies, music, news, user emotion plays an unexpected critical roles in the decision process. As the traditional model of user profile does not take into account the influence of the user emotion; the RS cannot recognize and capture the repetitively changing preferences of the user's like. Emotional status determination of the user is the main role of this algorithm, as it determines the state of the user's emotion according to three chosen colors by the user. In order to do so, it will analyses the color sequence by using the follow logic; if at least two of three chosen colors indicate the same emotion and this emotion becomes the current emotional state of the user [4-5].

The future is a viable opportunity to introduce a contextualized personalized and emotional RS with the ability to implement Multi-Agent System (MAS), sports among other domains. It makes such retrieval system necessary that along with the task of information gathering, could also involve in selective filtering as per the interests and emotions triggered by the information in the subject. For instance, today's hectic routine creates impediment for people in remaining up-to-date with respect to ones' social circle or world happenings [6]. It necessitates the embedding of users' intentions, their social networking habits and community trends into RS application. For accomplishing this purpose MAS intended to relay selective information to the subject is ought to be used [7].

The problem is the movie RSs need to capture exactly the customer's profile and movie features because the movie is a

complex domain, and emotion is a human interaction domain, so difficult to combining in the new recommender system. In this paper, we will apply matrix for integrating movie recommendation by hybrid approach, which consists of CBF and CF system with emotions detection algorithm and our algorithm. This system much better recommendations to users because it enables the users to understand the relation between their emotional states and the recommended movies.

This paper organized as follows: section 2 explains the background work of the movie RS and RSs based on emotion. Section 3 explained the recommendation approaches, human emotions, and combination the approaches. Section 4 demonstrates our algorithm to improving MRS based on emotion. In the section 5 presents some concluding remarks and points to future works.

II. Background and Related Work

In the last years, many RSs designed, and several RSs based on human emotions completed. This section presents some works that published by researchers.

According to [8], the CBF systems done each item by its related features. They learn a profile of the user's interests based on the features present in the items that the user has rated. The systems make recommendations by considering the description of the items that has been rated by the user and the depiction of items to be recommended. The recommendations can be made even if the system has received a small number of ratings, as the recommendations are based on product features. However, CBF systems are inadequate by the features that are explicitly related with the objects that they recommend [9].

By using MAS approach and taking advantage from some techniques of some other fields like Artificial Intelligence (AI) and Affective Computing (AC) this study Costa, H. and L. Macedo,2013 have been answered to the growing demand on RS. More precisely, it cope the privation of emotion-based RS, especially in relation to real-time textual information. While there is no study that related to this kind of the RS for Portuguese or even for English. This work will be focused in the development of a RS capable of filtering irrelevant and emotionless news to the user, by using a MAS approach [10].

Quijano-Sanchezat.al, 2011 have been issued a movie recommender system as an application in name Happy Movie (HM), for group of people which it is integrated with the social network face book (FB). They tried through HM to diminish a certain limitations in existing group RSs, like obtaining the user's profile or offering trading methods for users in order to reach a final agreement. The utilized method to make the group recommendation is based on three important features: personality, social trust and memory of past recommendations. Eventually, they simulate in a more realistic way the argumentation process followed by groups of people when deciding a joint activity [11].

Yang, and Yuan, 2010 presented a novel recommender service system to capture tourists' expectations (emphasizing on emotional needs) and to meet their satisfactions by recommending desirable attractions and SMEs in destination regions. Image, as a fundamental element and an operant resource, has been cultivated to be the uniform representation for tourists' expectations, destinations, and SMEs. In the system, the module constructs the images in three formats through the analysis of data; the interaction module and adaption module monitor the expected changes to images through the interactions between roles, and the unexpected changes caused by occasional environmental/social events respectively [12].

Menezes, and Tagmouti, introduced an Emotion-based Movie Recommender System (E-MRS) as a solution to this problem. The objective of E-MRS is to provide adapted and personalized suggestions to users using a combination of CF and content-based techniques. The recommendation is based on inferences about a user's emotions and preferences, as well as opinions of other similar users. In their paper discuss the system design and implementation, as well as its evaluation procedure. We believe that our system provides much better recommendation to users because it enables the users to understand the relation between their emotional states and the recommended movies [5].

Rajenderan, A, 2014 complete a project is to create a movie RS that uses human emotion as the basis of recommendation. The system will observe a user while they watch a portion of content, and then analyze the data that is their facial expressions and heart rate over time. With a sufficient database, their emotion data can be compared with those of other users, and recommendations can be provided based on reactions to the content [13], furthermore explains the landscape of actual and possible hybrid recommenders, and introduces a novel hybrid, a system that combines content boosted recommendation and CF to recommend restaurants [14].

Hong and Zhu, in their study proposed a novel method to measure users' preference on movie genres, and utilize Pearson Correlation Coefficient (PCC) to calculate the user similarity. For genre preference, regularization they used matrix factorization framework. Experimental results on Movie Lens data set demonstrate that the approach performs well. Their method can also be used to increase the genre variety of recommendations to some extent [15], moreover many articles for improving movie recommender system and emotion done, see [10, 16-20].

However many RSs for solving complexity of MRS was prepared, also they used hybrid approach. For developing MRS based on human emotions need to enhanced through new techniques or merging two or more techniques, in the next sections we explain and prepare a new hybrid approach to improving MRS.

III. Recommendation Approaches and Emotions

There are three major methods or approaches for RS: CF, CBF and Hybrid approach which combines the previous two methods. The first experiments in the RS area adopted pure CF approaches. Many researcher used CBF techniques, which described and utilized are borrowed from the field of information retrieval. However, CBF differs from information retrieval in the manner in which the interests of the users are represented. CBF tries to model the user's long-term interests instead of using a query of an information filtering system.

There are other systems that utilize content-based filtering to help users find information on the web. This section explains the RSs and emotions that used with MRS.

A. Recommendation System (RS)

RSs are information search tools that have been recently proposed to cope with the "information overload" problem, the typical state of the web user, of having too much information to make a decision or remain informed about a topic. In fact, users who are approaching a commerce website (e.g., Amazon), or a content website (e.g., cnet.com or visitfinland.com) for collecting information about a product or service, or simply a topic (e.g., Lapland) could be overwhelmed by the quantity of the relevant pages and ultimately the information displayed in these websites. In order to address this problem RS have been proposed.

The core computational task of an RS is to predict the subjective evaluation a user will give to an item. This prediction is computed using a number of predictive models that have a common characteristic, they exploit the evaluations/ratings provided by users for previously viewed or purchased items. Based on the particular prediction technique being employed, recommender systems have been classified into the following four main categories [9], Fig.1 shows a recommend system work.

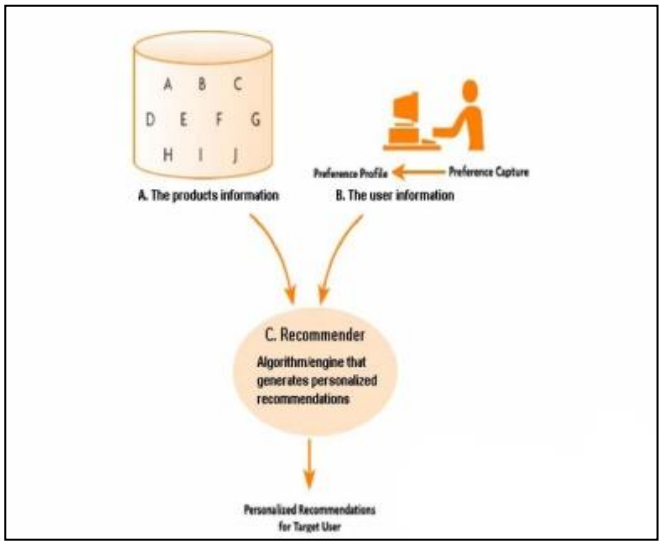


Fig. 1. Recommender System [21]

B. Objective from Recommender System

Prime objective of RS is to deliver to the user most relevant content as determined by ratings and previous consumption, in order to maximize positive user experience. Information overload is one of the main motivational factors behind this project the application of which could result in relevant content, thereby cementing the quality of recommendations. As opposed to 100-200 probably relevant recommendations , through the system under consideration we focus on delivering limited recommendations; though of higher quality reflects in our guarantee of the recommendations being more relevant [22].

C. Recommendation Approaches

There are many approaches that are used in recommender systems; Fig.2 shows the popular techniques of the RS.

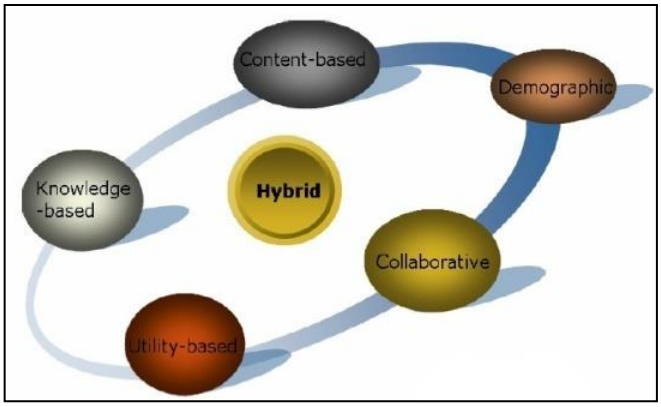


Fig. 2. Recommendation Techniques [9]

1) Content-Based Filtering (Item-Item Comparison)

CBF compares items against either user's interest profile or query derived from content. Strength of the technique over CF lies in recommendation delivery even when ratings received are lesser or none at all, as long as there is certain information pertaining to each item in the system. Each item however must be characterized in alignment with the features in user's profile. These descriptive features are either acquired or engineered. One paper introduced p-tango method, which uses both; content-based and collaborative filtering, by linear combination to create more effective filter than that produced through either method alone [4].

2) Collaborative Filtering (User-User Comparison)

It is also known as user-user contrast. One of the widely used recommendation methods is CF. Its basic idea is identification of likeminded users along with their cross-recommended items; items that likeminded users have liked, but the user on the receiving end has not consumed. Increasing popularity of this method could be attributed to the music industry where it is capitalized by such well-known websites as Pandora (www.pandora.com). One of the advantages it has is that it refrains from heavy calculations and so can produce highly accurate recommendations for a significant number of users in a timely manner. As the method draws on user-user similarity information to determine and pair likeminded users; users who are likely to agree rating of rate some specific content, it does not require any type of description regarding content as it only uses users' ratings for evaluation of their degree of agreement. User-user similarity information includes rating given by both users. Rating differences that determine user like-mindedness are stored in a database to intercept similar users when generating recommendations.

In [23] suggests a good structure for the similarity values in a table called a user-user matrix. The recommendations can now be produced by few variations of methods. For example find the most often rated or highest rated items on average from the users N closest neighbors that he has not seen or rated himself.

What we however decided to do was a method inspired by the Group Lens team from [8]. By using the top N nearest neighbor's ratings and similarity values, we can estimate a predicted rating by using a weighted average of the neighbor ratings. The highest predicted ratings will then give us a good idea of the user's interest and have shown to produce high quality recommendations. In-spite of the method's growing popularity there are two main problems that are generally associated with collaborative methods. The first issue is caused by new users in the systems which have not submitted any ratings. The system is therefore unable to find any qualified user-neighbors' and thereby also unable to predict any ratings based on that. This problem is called Cold-start problem. A simple solution to the problem is not to offer recommendations to users that have not submitted any ratings or perhaps fewer than some minimum number of ratings. The other perhaps more relevant problem is called the first-ratter problem. This problem is caused by new items in the systems that understandably have not received any ratings.

3) Combination Recommendation Approaches

Utilizing different methods are often related to the different recommendation issues. These problems are caused by four main instances that we try to avoid them by combining collaborative and content-based techniques to compensate each other downsides. The main causes are new items, new users that understandably have not submitted or received any ratings. Sparsely due to large data sets and little overlapping between users rating and finally averaging effect e.g. due to long time content consumption and poor system design [16], as shown in Fig 3.

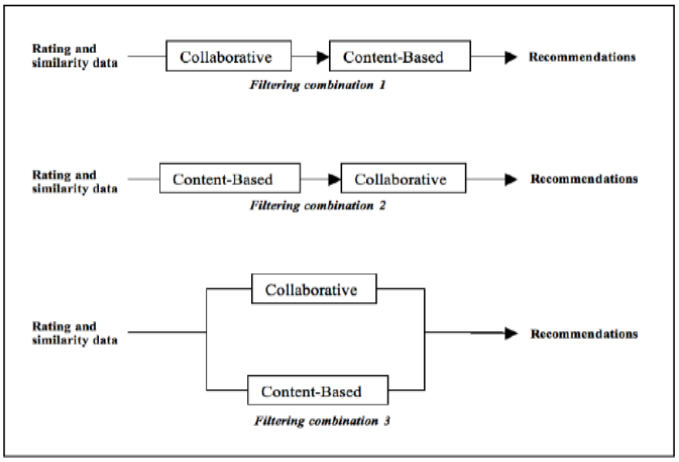


Fig. 3. Different Ways of Combining CF and CBF [16]

D. Emotion States

Various aspects of emotion are measurable including self-reported individuals' feelings, autonomic system reactions, neurological changes, and bodily actions inclusive of facial movements. Others nevertheless continue to argue presence of universal fundamental emotions; emotions that could be universally recognized, and their existence due to evolutionary pressure. For instance, cognitive processes and autonomic changes during fear induce flight response in human preparing them to run away from the threat or danger. A scientific evidence is highlighting five different emotions at least (anger, fear, disgust, sadness, happiness) being noticeably different considering the activation of different combinations in the brain supports the above phenomenon [24].

Emotion is a state usually triggered by any significant event characterizing importance of some level to the subject or user. It normally includes (a) conscious mental state having discernable quality of feeling directed towards any object, (b) some kind of bodily perturbation, (c) identifiable expressions on the face, tonality of the voice, and gesture (d) readiness to indulge in certain types of actions[25]. Emotion can be identified as a series of changes in the state, the changes that are way inter-dependent and synchronized in a way in response to evaluations to which relevance of internal or external stimulus are subjected.

People or viewers watch and experience movies everyday with affective response. They depict joy when watching a comedy and sadness while watching romantic movie late into the night. Another way to prove this is when somebody is in love; they would like to watch romantic movies that generally follow happy endings. However, when that same person is sad, they might watch fast-paced action flick as it could improve their feelings through subversion of their sadness that is triggered by their active involvement in the fast-paced movie to understand what is happening. Fast-paced movies usually overcome feelings of viewers; usually negative, as they indulge in decoding the scenarios unraveling in front of them in speed.

Another approach that could be utilized is learning the rules with the help of training. In this scenario, the user could be asked and required to answer several questions aided by the use of questionnaire (movie-emotion). The answer or user response to the questionnaire would indicate the preferences; with respect to movie, of the subject as per his emotions. It could be intercepted through such questions as the kind of movies one would like to watch when in the state of sadness, happiness etcetera. This method is usually chosen to be used for the creation and development of basic rules pertaining to the system focused in this assignment. Quality and quantity of the developed rules is imperative to be improved over a period with increasing number of users enrolling up for the system.

IV. Methodology and System Framework

This section explained the progress of a new movie recommender system based human emotions, which consist of CF and CBF after mixing emotion detection algorithm with our method to improvement the system.

The system framework consists of five phases as shown in Fig.4, in the first phase the users should be make a registration and chose three colors to selecting emotion state. In the second phase the system calculate initial rating and allow users to rating the movies from 1 to 5, in the third phase system calculate the similarity between users and items, in the fourth phase the system predicting the user's rating through hybrid RS, in the last phase the system sort the list of movies that recommended for the target user.

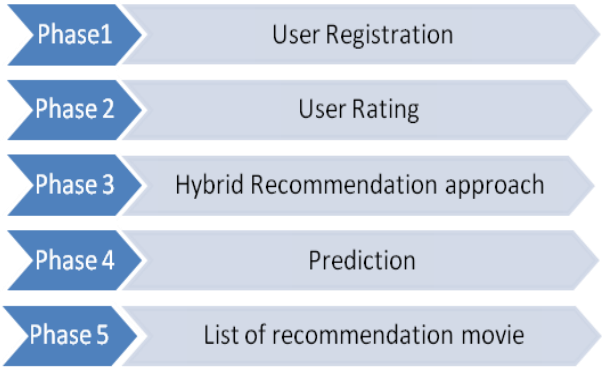


Fig. 4. System Framework

Phase 1: The users must start working in a special account in the system and enter the required information upon it such (id, user name, password, email, choosing three colors), then the three colors converted to emotion state by the emotion detection algorithm as shown in Fig.5.

```
If choice1 in (Y,LO,B,G) and choice2 in (Y,LO,B,G) and
choice3 in (Y,LO,B,G) then
State = joy
Else if choice1 =LR and choice2 =LR and choice3 = LR then
State = love
.
.
.
Else if (choice1 and choice2 in (Y,LO,B,G,LR) and choice3 not
in (Y,LO,B,G,LR) ) OR
(choice2 and choice3 in (Y,LO,B,G,LR) and choice1 not in
(Y,LO,B,G,LR)) OR
(choice3 and choice1 in (Y,LO,B,G,LR) and choice2 not in
(Y,LO,B,G,LR)) then
State = joy and love
Else if (choice1 and choice2 in (DR,BR,GR,BK) AND choice3
NOT IN (DR,BR,GR,BK) ) OR (choice2 and choice3 in
(DR,BR,GR,BK) AND choice1 NOT IN (DR,BR,GR,BK) )
OR (choice3 and choice1 in (DR,BR,GR,BK) AND choice2
NOT IN (DR,BR,GR,BK) ) THEN
State = anger and Sadness
End if
```

Fig. 5. Emotion Detection Algorithm

Phase 2: After registration the users can evaluate the movies by rating movies from 1 to 5. The users after seen the part or the whole movie will decide the movie according to his/her like, then rated between 1 to 5, as shown in Table 1.

TABLE I. USER–ITEM MATRIX RATING

	Item 1	Item 2	Item 3	Item 4	Item 5	Item 6	Item 7
User 1	4	3	0	2.5	3.5	5	0
User 2	5	4	0	4.5	2.5	0	0
User 3	0	2.5	3	0	1	0	1.5
User 4	0	0	0	2.5	0	3.5	4
User 5	1	0	5	0	0	0	0

Phase 3: This phase calculates a new hybrid recommender system that consists of CF and CBF with human emotions and our algorithm as shown in Fig.6.

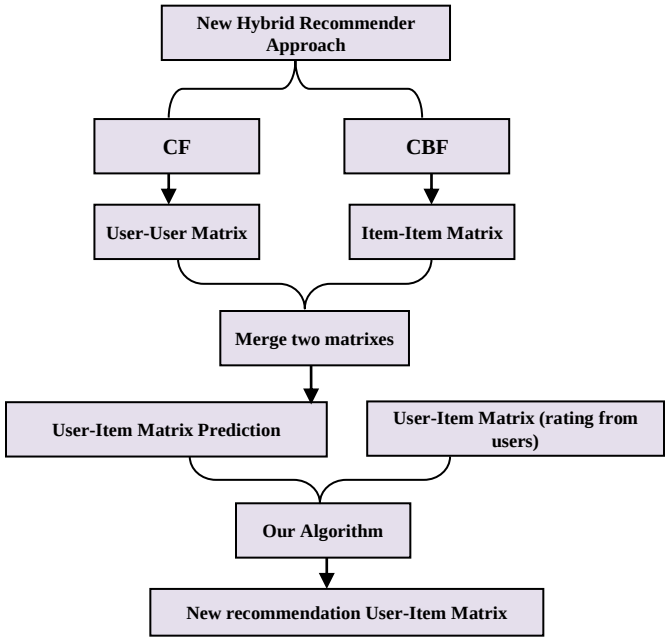


Fig. 6. New Hybrid Recommender System

In the following explained and calculated the steps of a new hybrid recommender system:

A. Calculation User-User Matrix (CF)

CF is a recommendation method used to finding similarity between users, as explained in section 3, we can find similarity between users through equation (1).

$$\text{Sim(user i),(user j)} = \frac{(n*mr)-\sum \text{abs}(r\,i(X)-r\,j(X))}{n*mr} \quad (1)$$

n: number of commonly rated items
mr: maximum rating value (5.0)
ri(x): Rating for item x from user(i), user(i) rated according to emotion state.
rj(x): Rating for item x from user(j), user(i) rated according to emotion state.

Example 1: Let's assume that two users in the system have submitted a few ratings, and from those ratings we want to compare how 'likeminded' these two users are. From all the ratings they have submitted only the following 5 items are overlapping, meaning items that both have rated, as shown in Table 2.

TABLE II. USER RATING

	User 1	User 2	Offset
Item 1	4.5	2	2.5
Item 2	3	2	1
Item 3	2.5	5	2.5
Item 4	2	5	3
Item 5	1.5	4.5	3
Total			12

After applied equation (1), the similarity between user1 and user2 is 52% as calculated the following:

$$\text{Sim(user1),(user2)} = \frac{5*5-12}{5*5} = \frac{25-12}{25} = 0.52 = 52\ \%$$

The similarity between users to get to the best result, and so when the user does a search for movies who likes to watch. To get the best result for users, we arrange movies by selected colors and the similarity between users.

Example 2: Let's assume that five users make ratings for five items, after finding the similarity between users by using equation (1), for easy present we use matrix for showing similarity between users as shown in Table 3.

TABLE III. USER-USER SIMILARITY MATRIX

	User1	User2	User3	User4	User5
User1	1	0.52	0.36	0.44	0.5
User2		1	0.58	0.67	0.32
User3			1	0.65	0.83
User4				1	0.78
User5					1

Moreover Fig 7 shows the php code for finding similarity user-user.

```
<?php
$host = "localhost"; //server
$db = "mad"; //database name
$user = "root"; //dabases user name
$pwd = ""; //password
$link = mysqli_connect($host, $user, $pwd, $db);
$query = "select user1,user2,(count(title)*5 -(sum(abs(user1-
user2))))/(count(title)*5)*100 as sim from sime_user";
$result = mysql_query($query);
$results = mysqli_query($link, $query);
if(mysqli_num_rows($results) >= 1)
{
      $output = "";
      while($row = mysqli_fetch_array($results))
      {
   $output .= "user 1: " . $row['user1'] . "<br />";
        $output .= "user 2: " . $row['user2'] . "<br />";
              $output .= "sime user 1-2: " . $row['sim'] . "<br />";
              }
      echo $output;
}
$query1 = "select (count(title)*5 -(sum(abs(user1-
user3))))/(count(title)*5)*100 as siim from sime_user";
$result = mysql_query($query1);
$results = mysqli_query($link, $query1);
if(mysqli_num_rows($results) >= 1)
{
      $output = "";
      while($row = mysqli_fetch_array($results))
      {
              $output .= "sime user 1-3: " . $row['siim'] . "<br />";
      }
      echo $output;
}
$query2 = "select (count(title)*5 -(sum(abs(user1-
user4))))/(count(title)*5)*100 as simm from sime_user";
$result = mysql_query($query2);
$results = mysqli_query($link, $query2);
if(mysqli_num_rows($results) >= 1)
{
      $output = "";
      while($row = mysqli_fetch_array($results))
      {
      $output .= "sime user 1-4: " . $row['simm'] . "<br />";
      }
      echo $output;
}
```

Fig. 7. php code for similarity user-user

B. Calculation Item-Item Matrix (CBF)

CBF is a recommendation method used to finding similarity between items, as explained in section 3, Calculating similarity between item-item should be defined item's profile; this attributes for a movie selected (title, genre, colors, country, and cast) as shown in Table 4.

TABLE IV. AN EXAMPLE OF AN ITEM-PROFILE FEATURING: GALDIATOR (2000)

Title	Gladiator (2000)
Genre	Action, Adventure, Drama
Colors	Black, red
Country	USA
Cast	Russell Crowe, Joaquin Phoenix, Connie Nielsen, Oliver Reed, Richard Harris, Derek Jacobi, Djimon Hounsou, David Schofield, John Shrapnel, Tomas Arana, Ralf Moeller, Spencer Treat Clark, David Hemmings, Tommy Flanagan,... and so on.

The attributes listed in Table 4 is a typical item-profile which can used to compare how much alike items are. But the main concern in this field should be how selected the right attributes and which attributes describe the content in the best possible way. Assuming that we have two items with a number of attributes, each similarity between them is calculated using the following equation (2).

$$\text{Sim(item(i), item(j))} = \sum wx * ax(i,j)/bx(i,j) \qquad (2)$$

wx: Weight-factor for attributes of type x
ax: Number of common, type x attributes from item i and j
bx: Lower number of type x attributes from item i and j

As previously mentioned, attributes can give different amount of information. For example, the fact an item was produced in USA or that it was produced in the year 2004 does not really tell us anything about the item's content. However, if the same director for example produced the item there is a considerable possibility that the items will share some

similarities. We used linear regression method on 100 randomly selected ratings to find the optimal weight-factor that would produce the minimum error in the rating prediction. The result of the regression method is the following weights:

W1: Genre 30.08
W2: Emotion 41.16
W3: Country 10.54
W4: Cast 9.08
W5:9.14

Example 3: let's assume that we have the following two items that we want to compare: "The Terminator (1984)" and "Terminator II (1991)" that presumably should give a relatively high similarity value since they have many of the same characteristics. See Table 5.

TABLE V. PROPORTION OF SIMILAR ATTRIBUTES – EXAMPLE

Title	The Terminator (1984)	The Terminator II (1991)	Common
Genre	Action, Sci-Fi, Thriller	Action, Sci-Fi, Thriller, Adventure	3/4
colors	Black, red	Black, red	2/2
Cast	Arnold Schwarzenegger, Michael Biehn, Linda Hamilton	Arnold Schwarzenegger, Michael Biehn, Linda Hamilton	3/15
Year	1984	1991	0/1
Country	USA	USA	1/1

After applying equation (2) on example 3, the similarity between the two movies is 82.48% as calculated below:

$$\text{Sim} = 0.3008 * \frac{3}{4} + 0.4116 * \frac{2}{2} + 0.908 * \frac{3}{15} + 0.1054 * \frac{1}{1} + 0.0914 * \frac{0}{1}$$

Sim= 92.42%

After calculated the similarity value for each item pair updated the value and the total number of common attributes in the item-item-matrix, N nearest neighbors with a minimum threshold of 5 common attributes that the neighbor has to have in common with the neighbor in order for it to qualify and be used for the prediction. The threshold limit was found by a few simple experiments and then evaluating the mean absolute error between the prediction and the real rating. The size of the threshold controls a balance between the quality of the neighbors and the amount of neighbors we use for the prediction. If a threshold is chosen too low (e.g. common > 2) the prediction may use item-neighbors that have high similarity value only because of bad/short descriptions and are irrelevant in relation to other items that have more attributes in common but still manage to get a lower similarity value. However if the threshold is chosen too high (e.g. common > 40) the prediction will possibly be made using too few neighbors or perhaps even no neighbors qualify which means that we are unable to make the prediction, at least from the item-CF side.

Example 4: Let's assume that seven movies according to item profiles by using equation (2), we found similarity between items, and using matrix for easy presentation similarity between movies, see Table 6.

TABLE VI. ITEM-ITEM SIMILARITY MATRIX

	Item 1	Item 2	Item 3	Item 4	Item 5	Item 6	Item 7
Item 1	1	0.48	0.84	0.74	0.23	0.69	0.95
Item 2		1	0.98	0.19	0.56	0.61	0.87
Item 3			1	0.90	0.54	0.55	0.71
Item 4				1	0.72	0.43	0.70
Item 5					1	0.51	0.63
Item 6						1	0.81
Item 7							1

Above example shows how the Item-Item-Matrix is constructed with the similarity values between the relative items. Example above each similarity value in the subscript style represents the number of common attributes used for each calculation. Fig 8 shows the php code for finding similarity between movies.

```
<?php
$con=mysqli_connect("localhost","root","","mad");
// Check connection
if (mysqli_connect_errno()) {
  echo "Failed to connect to MySQL: " . mysqli_connect_error();
}
$result = mysqli_query($con,"select * from item");
echo "<table border='1'>
<tr>
<th>title</th>
<th>gener</th>
<th>emotion</th>
<th>cast</th>
<th>Country</th>
</tr>";
while($row = mysqli_fetch_array($result)) {
        echo "<tr>";
        echo "<td>" . $row['title'] . "</td>";
        echo "<td>" . $row['gener'] . "</td>";
        echo "<td>" . $row['emotion'] . "</td>";
        echo "<td>" . $row['cast'] . "</td>";
        echo "<td>" . $row['country'] . "</td>";
        echo "</tr>";
      }
      echo "</table>";
      $result = mysqli_query($con,"select (0.25*
(count(DISTINCt(gener)))/(count(gener) ))+(0.25*
(count(DISTINCt(title)))/(count(title) ))+(0.5*
(count(DISTINCt(emotion)))/(count(emotion) ))+(0.25*
(count(DISTINCt(country)))/(count(country) )) as sim from item");
      echo "<table border='1'>
      <tr>
      <th>similarity item1 and item2</th>
      </tr>";
      while($row = mysqli_fetch_array($result)) {
        echo "<tr>";
        echo "<td>" . $row['sim'] . "</td>" ;
        }
        echo "</table>";
        echo "</table>";
        ?>
```

Appendix3: Prediction Code

```
<?php
$searchTerm = trim($_GET['keyname']);
if($searchTerm == "")
{
        echo "Enter name you are searching for.";
        exit();
}
$host = "localhost";
$db = "mad";
$user = "root";
$pwd = "";
$link = mysqli_connect($host, $user, $pwd, $db);
$query = "SELECT `user id`, `item id`, `rating`, `emotion` FROM
`rating` WHERE name LIKE '%$searchTerm%' and rating>=3 and
emotion ="action"";
$results = mysqli_query($link, $query);
if(mysqli_num_rows($results) >= 1)
{
$output = "";
while($row = mysqli_fetch_array($results))
{
$output .= "<video width='320' height='240' controls ><source
src='rating/". $row['vid'] .".mp4' type='video/mp4'> </video>"   . "<br
/>";
echo $output;
}else
echo "There was no matching record for the name " . $searchTerm;
?>
```

Fig. 8. php code for item-item similarity

C. Calculation Prediction (User-Item Matrix)

For finding prediction, items for each user should be calculated prediction for items per users by using equation (3), and then we can get user-item matrix as presents in Table 7.

$$P\ i,u = \frac{\sum S\ i,n * R\ u,n}{\sum S\ i,n} \quad (3)$$

si,n: Similarity value between target item i and neighbor n

Ru,n: Rating from neighbor n

Example 5: for predicting rating user 1 to item 1 we use equation 3, through the neighbor's similarity user's ratings.

Sim user1 with neighbours: 0.52, 0.36, 0.44, 0.5

Rating user1 for items: 5,0,0,1

$$P1 = \frac{0.52*5+ 0.36*0+ 0.44*4+ 0.5*1}{0.52+ 0.36+ 0.44+ 0.5} = 2.67$$

Sim item1 with neighbours: 0.48, 0.84, 0.74, 0.23, 0.69, 0.95.

Rating: 3, 0, 2.5, 3.5, 5,0

$$P2 = \frac{0.48*3 + 0.84*0 + 0.74*2+0.23*5+0.69*5+0.95*0}{0.48 + 0.84 + 0.74+0.23+0.69+0.95} = 1.91$$

The final prediction (P) is prediction between P1 and P2 as calculated the following:

$$P = \frac{P1 * N1 + P2 * N2}{N1 + N2} = \frac{2.67 *4 +1.91 * 6}{4 + 6} = 2.21$$

TABLE VII. USER – ITEM PREDICTION

	Item 1	Item 2	Item 3	Item 4	Item 5	Item 6	Item 7
User 1	2.21	2.7	2.55	3.5	3.05	3.8	3.2
User 2	2.8	3.9	3.5	4.5	2.75	2	4
User 3	4	4.3	2.3	3.5	2.2	3.5	2.35
User 4	3.2	4.1	2.71	1	3.1	3.5	2.5
User 5	1.3	4.5	5	1.2	4.1	2.6	3.3

D. Calculate Our Approach New User-Item Matrixes and List Of Prediction

After rating movies from users (see Table 1), and finding prediction for movie's rating from users (see Table 7), we get two user-item matrixes, our approach is combining these two matrixes after apply a condition. The condition is replace ratings table 1 to table 7 with remain 1 and 5, because number 1 meaning the user absolutely dislike the movie and number 5 meaning the user absolutely like the movie, so no need to find predictions, but ratings between 1 and 5 may changed after more rating from other users that profiles near the target user, as shown in Table 8.

TABLE VIII. USER-ITEM OUR ALGORITHM

	Item 1	Item 2	Item 3	Item 4	Item 5	Item 6	Item 7
User 1	2.21	2.7	2.55	3.5	3.05	**5**	3.2
User 2	**5**	3.9	3.5	4.5	2.75	2	4
User 3	4	4.3	2.3	3.5	**1**	3.5	2.35
User 4	3.2	4.1	2.71	1	3.1	3.5	2.5
User 5	**1**	4.5	**5**	1.2	4.1	2.6	3.3

In the final step the system list the top movies for target user, the system predict best movies for the target user that like watch it.

E. Review System

To review the system, these figures below show the user interface in Fig.9, and system framework in Fig.10.

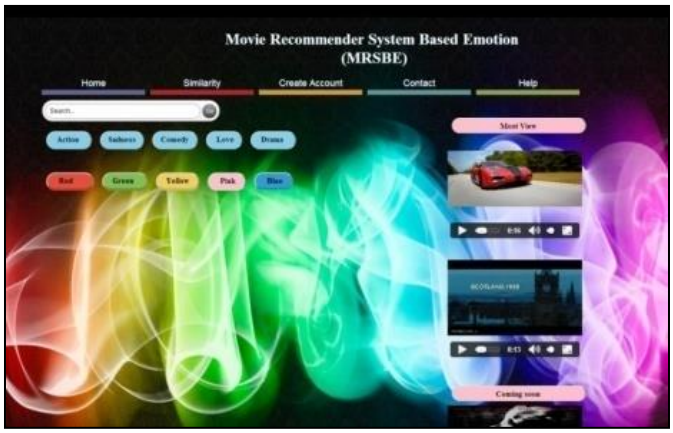


Fig. 9. Interface New Movie Recommender System

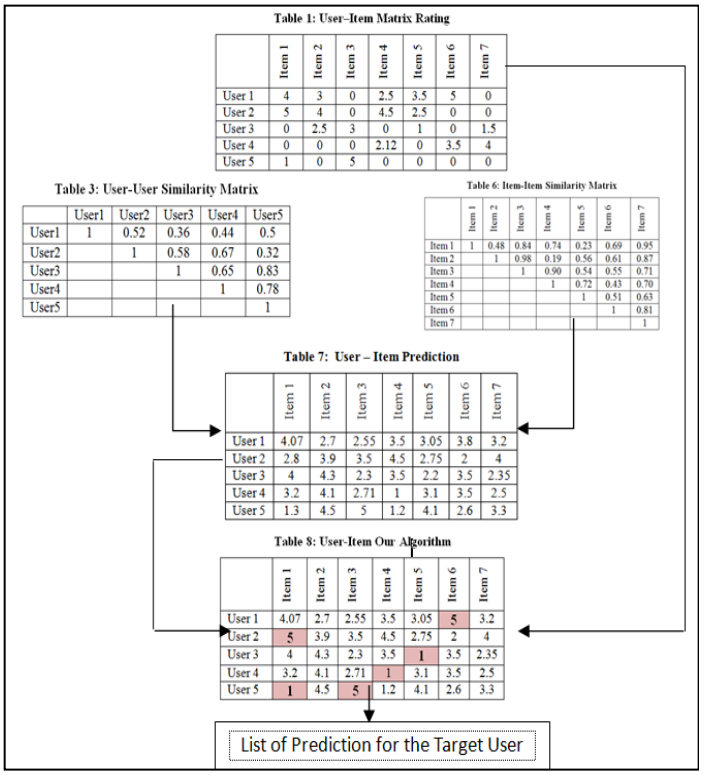


Fig. 10. System Framework

V. Conclusion and Future Work

In this paper we designed a new web recommender system for movies based on emotion. The movies are a complex object and emotions are a human interaction, which is difficult combining together. In this paper, we applied matrix for integrating movie recommendation by hybrid approach, which consists of CBF and CF system with emotions detection algorithm and our algorithm. Furthermore our algorithm calculated the user rating 1 and 5 because the users absolutely liked or disliked the movies this system much better recommendations to users because it enables the users to understand the relation between their emotional states and the recommended. We recommend the researchers to improve this idea through: 1) Extracting the movies to finding most using colors by system. 2) Using more than two recommendation techniques to getting best capture of the movies. 3) Using more than three colors to finding human emotions. 4) Design a new algorithm to solving the movie recommender system.

References

[1] Ricci, F., Rokach, L., &Shapira, B. (2011). "Introduction to recommender systems handbook": Springer.

[2] Eyjolfsdottir, E. A., Tilak, G., &Li, N.(2010) "MovieGEN: A Movie Recommendation System", UC Santa Barbara: Technical Report.

[3] Shani, G. &Gunawardana, A. (2011). "Evaluating recommendation systems", in Recommender systems handbook, ed: Springer, 2011, pp. 257-297.

[4] Pazzani, M. J. &Billsus, D. (2007). "Content-based recommendation systems", in The adaptive web, ed: Springer, 2007, pp. 325-341.

[5] Ho, A. T., Menezes, I. L., &Tagmouti, Y. (2006) "E-MRS: Emotion-based movie recommender system", in Proceedings of IADIS e-Commerce Conference. USA: University of Washington Both-ell, pp. 1-8, 2006.

[6] Joly, A., Maret, P., &Daigremont, J. (2010) "Enterprise Contextual Notifier, Contextual Tag Clouds towards more Relevant Awareness", in Proceedings of the ACM Conference on Computer Supported Cooperative Work, pp. 531-532, 2010.

[7] Costa, H.(2012) "A Multiagent System Approach for Emotion-based Recommender Systems", PhD proposal, University of Coimbra, Coimbra, Portugal.

[8] Good, N., Schafer, J. B., Konstan, J. A., Borchers, A., Sarwar, B., Herlocker, J., &Riedl, J. (1999) "Combining collaborative filtering with personal agents for better recommendations", in AAAI/IAAI, pp. 439-446, 1999.

[9] Perdew, J. P., Burke, K., &Ernzerhof, M.(1996) "Generalized gradient approximation made simple", Physical review letters, vol. 77, p. 3865.

[10] Costa, H. &Macedo, L. (2013). "Emotion-Based Recommender System for Overcoming the Problem of Information Overload", in Highlights on Practical Applications of Agents and Multi-Agent Systems, ed: Springer, 2013, pp. 178-189.

[11] Quijano-Sanchez, L., Recio-Garcia, J. A., &Diaz-Agudo, B. (2011) "Happymovie: A facebook application for recommending movies to groups", in Tools with Artificial Intelligence (ICTAI), 2011 23rd IEEE International Conference on, pp. 239-244, 2011.

[12] Yang, C.-Y. &Yuan, S.-T.(2010) "Color Imagery for Destination Recommendation in Regional Tourism".

[13] Rajenderan, A.(2014) "An Affective Movie Recommendation System".

[14] Melville, P., Mooney, R. J., &Nagarajan, R. (2002) "Content-boosted collaborative filtering for improved recommendations", in AAAI/IAAI, pp. 187-192, 2002.

[15] Nie, D., Hong, L., &Zhu, T. (2013) "Movie Recommendation Using Unrated Data", in Machine Learning and Applications (ICMLA), 2013 12th International Conference on, pp. 344-347, 2013.

[16] Peleja, F., Dias, P., Martins, F., &Magalhães, J.(2013) "A recommender system for the TV on the web: integrating unrated reviews and movie ratings", Multimedia systems, vol. 19, pp. 543-558.

[17] Kim, M. &Park, S. O.(2013) "Group affinity based social trust model for an intelligent movie recommender system", Multimedia tools and applications, vol. 64, pp. 505-516.

[18] Liang, T., Wu, S., &Cao, D. (2012). "Improved Collaborative Filtering Method Applied in Movie Recommender System", in Emerging Computation and Information teChnologies for Education, ed: Springer, 2012, pp. 427-432.

[19] Jung, K.-Y., Park, D.-H., &Lee, J.-H. (2004). "Personalized movie recommender system through hybrid 2-way filtering with extracted information", in Flexible Query Answering Systems, ed: Springer, 2004, pp. 473-486.

[20] Fernández-Tobías, I., Cantador, I., &Plaza, L. (2013). "An Emotion Dimensional Model Based on Social Tags: Crossing Folksonomies and Enhancing Recommendations", in E-Commerce and Web Technologies, ed: Springer, 2013, pp. 88-100.

[21] Lichtenberg, A. J. &Lieberman, M. A.(1983) "Regular and stochastic motion", Research supported by the US Department of Energy, US Navy, and NSF. New York, Springer-Verlag (Applied Mathematical Sciences. Volume 38), 1983, 516 p., vol. 38.

[22] Gaczynska, M., Rock, K. L., &Goldberg, A. L.(1993) "γ-Interferon and expression of MHC genes regulate peptide hydrolysis by proteasomes".

[23] Chinwalla, A. T., Cook, L. L., Delehaunty, K. D., Fewell, G. A., Fulton, L. A., Fulton, R. S., Graves, T. A., Hillier, L. W., Mardis, E. R., &McPherson, J. D.(2002) "Initial sequencing and comparative analysis of the mouse genome", Nature, vol. 420, pp. 520-562.

[24] Thamm, R.(1992) "Social structure and emotion", Sociological Perspectives, vol. 35, pp. 649-671.

[25] Artymiuk, P., Blake, C., Grace, D., Oatley, S., Phillips, D., &Sternberg, M.(1979) "Crystallographic studies of the dynamic properties of lysozyme".

Service Design for Developing Multimodal Human-Computer Interaction for Smart Tvs

Sheng-Ming Wang
Department of Interaction Design
National Taipei University of Technology
Taipei, TAIWAN

Cheih-Ju Huang
Ph.D. Program, College of Design
National Taipei University of Technology
Taipei, TAIWAN

***Abstract*—A Smart TV integrates Internet and Web features into a TV, as well convergence between computer and TV and can utilize as a computer. Smart TV devices facilitate the curation of content by combining Internet-based information with content from TV providers. Many techniques, such as those that focus on speech, gestures, and eye movement, have been used to develop various human computer interfaces for Smart TVs. However, as suggested by several researchers, user scenarios and user experiences should be incorporated with development techniques to meet user demands on Smart TVs. Thus, this study applies the service design approach for scenario planning and user experience analysis of multimodal interaction development for Smart TVs. This research begins with the service design process and derives the Quality Function Deployment matrix (QFD Matrix) for initial decision-making. Analytical Hierarchy Process (AHP) is then applied to evaluate the priority and relevance of features proposed in the QFD Matrix. Research results show the service design approach is an efficient way for an interdisciplinary team to communicate. The proposed two-stage decision-making processes provide qualitatively analyze and quantitatively measure the priority and relevance of features derived from the service design process. The technique team can then develop prototypes that facilitate multimodal human-computer interaction on Smart TVs.**

Keywords*—*Smart TV; Service Design; Human-Computer Interaction; Quality Function Deployment; Analytical Hierarchy Process

I. INTRODUCTION

A smart TV is either a TV with integrated Internet capabilities or a television with a set-top box that offers advanced computing abilities and connectivity than a typical TV. Smart TVs may be considered a TV that also hascomputer operating system that often allows users to install and run advanced applications on a specific platform.

A Smart TVcan broadcast broadband web content[1]. It has the potential to seamlessly integrate the strengths of TV broadcasting and broadband network services[2]. Smart TVs currently provide access to the Internet and legacy web services, and specify which content services are immediately coupled to broadcast content that is rendered on the terminal device[3].

Although a Smart TV attempts to serve audiences through its innovative services, a number of questions remain about the mechanism delivering services to different users *via* the same platform. Additionally, Smart TVs require innovative human-computer interactions to provide enhanced services and fulfill user requirements[4]. Differing from a conventional TV with a remote control, new Smart TV features, such as web search, social networking, multi-user, personalized services and applications development, require innovative "natural" human-computer interactions. Using keystrokes on remote controls, touching the TV screen, or using the touch panel on smart handheld devices are inconvenient and impose limitations on users. Some user groups, such as the disabled or elderly in particular face problems when using these services[5]. The availability of accessible user interfaces that can adapt to the specific needs of users with impairments is very limited. Notably, no method automatically adapts to multimodal interactions, such that they cannot automatically fit the requirements of users with different impairments[5].

To improve the multimodal human-computer interaction of Smart TVs, one must bring together technicians and designers inter-disciplinary integration to generate a comprehensive roadmap for development and identify the future requirements for Smart TVs[4, 5]. This research uses the service design approach to organize a cross-discipline professional team, including of computer science and interaction design professionals, to evaluate the features of human-computer interaction mechanisms of Smart TVs. The principles of service design were implemented in scenario planning[6]. The quality function deployment (QFD) matrix, a qualitative approach, systematically assesses the correlation between user requirements and technical features [7]. Finally, the analytical hierarchy process (AHP) synthesizes the features in the QFD matrix and ranks alternatives. Therefore, both qualitative and quantitative criteria can be weighted and prioritized using informed judgments[8, 9].

The remainder of this paper is organized as follows. Related work is analyzed in Section II. Section III discusses the design mechanism and proposed methodology. Implementation results of QFD and AHP are then discussed in Section IV. Conclusions and future works are presented in Section V.

II. RELATED WORK

According to [10, 11], "service interfaces are designed for intangible products that are, from the customer's point of view, useful, profitable and desirable, while they are effective, efficient and different for the provider." The method for making this process integral and holistic is to incorporate the particular visions of all stakeholders, including users,

designers, investors, researchers, technicians, policy makers, consultants and competitors [12]. Moggridge asserted that "service design is the design of intangible experiences that reach people through many different touch-points"[11]. That is, service design is a process of continual updates based on the responses of users who are observed and monitored.

Any application of service design to the multimodal interaction development of Smart TV must consider aspects of both product design and interface design. As pointed out by Obrenovic and Starcevic[13], multimodal interfaces move the balance of interactions closer to the human and offer expressive, transparent, efficient, and robust human-computer interactions. In human-computer interactions, the term modality typically refers to the five human senses—sight, hearing, touch, smell, and taste. Oviatt[14] offered a more practical definition, stating that multimodal systems coordinate the processing of combined natural input modalities, such as speech, touch, hand gestures, vision, head and body movements, with multimedia system output. Thus, by applying a service design approach to the multimodal interaction of Smart TVs, this work follows Oviatt's definition [14] and focuses on applications of speech, touch, hand gestures and visualizing. Moreover, this work follows some features and characteristics of service design that were summarized by Blomkvist and Holmlid[15, 16], including the following.

1) Assessing services from a holistic and detailed point of view.

2) Considering both artifacts and experiences.

3) Making services tangible and visible via visualizations.

In addition to the service design approach, the QFD matrix and AHP are also utilized simultaneously to systematically identify the criteria derived from service design scenario planning, and to weight and prioritize criteria.

As proposed by many researchers, the QFD matrix transforms customer requirements (CRs) into technical requirements. Originally a quality improvement tool, the QFD matrix has been widely used to develop new products [7, 17, 18]. For effective product design, a design team must be cognizant of what they are designing and what users will expect. The QFD matrix is a systematic design approach based on an in-depth awareness of customer desires, coupled with integrated corporate functional groups. The QFD matrix translates customer desires into design characteristics for each stage of product development. The ultimate goal of is to translate often subjective criteria into objective criteria that can be quantified and measured and which can then be used to design and manufacture the product. According to Akao[17], weighting customer requirements is a critical step in building a QFD matrix. The simples tmethod is to ask an expert panel to apply a point scale, such as 1–5 or 1–9 and score each CR. However, this simple system has two weaknesses: it does not prioritize customer requirements; and weights are subjective and depend on panel consensus.

To overcome these weaknesses, several researchers and practitioners have advocated using the AHP to weight CRs. The AHP is a structured technique for dealing with complex decisions[19–21]. Conditions of uncertainty arise from subjective information (presented as quantitative and qualitative values) used in decision-making processes. This uncertainty is based on incomplete decision knowledge about the properties of an object, insufficient confidence in the accuracy of expert judgments, knowledge inconsistencies, and information fuzziness[22]. Therefore, implementation of a decision-making problem under uncertainty requires a comparison of factors lacking quantitative characteristics or a simultaneous comparison of quantitative and qualitative characteristics. In the capacity of tool for such problems decision heuristic methods based on expert judgments may be used in addition to the AHP[9].

The AHP enables groups of people to interact and focus on a certain problem, modify their judgments and, as a result, combine group judgments in accordance with the main criteria[23]. Applying the AHP to weight CRs in a QFD matrix provides a rational framework for structuring a decision problem. The combined AHP-QFD approach can quantify CRs and elements, relate those elements to overall CR goals and evaluate alternative solutions. The combined AHP-QFD approach has been used successfully to assess customer needs based on a multiple-choice decision analysis[24]. Gupta *et al*. [25] reviewed uses of the QFD-AHP to evaluate and select methodology for an innovative product design concept. The methodology combining QFD-AHP was mainly used as a multi-criteria decision method for evaluating user requirements. By considering the multimodal interaction requirements of Smart TVs and characteristics of service design, this work uses this methodology to evaluate the multimodal HCI design and development of Smart TVs.

III. METHODOLOGY

This work integrates design thinking with technology development process for developing a multimodal Human Computer Interface (HCI) for Smart TVs. Figure 1 shows the comprehensive structure of the integration of design thinking concept with technology development an inter-disciplinary approach.

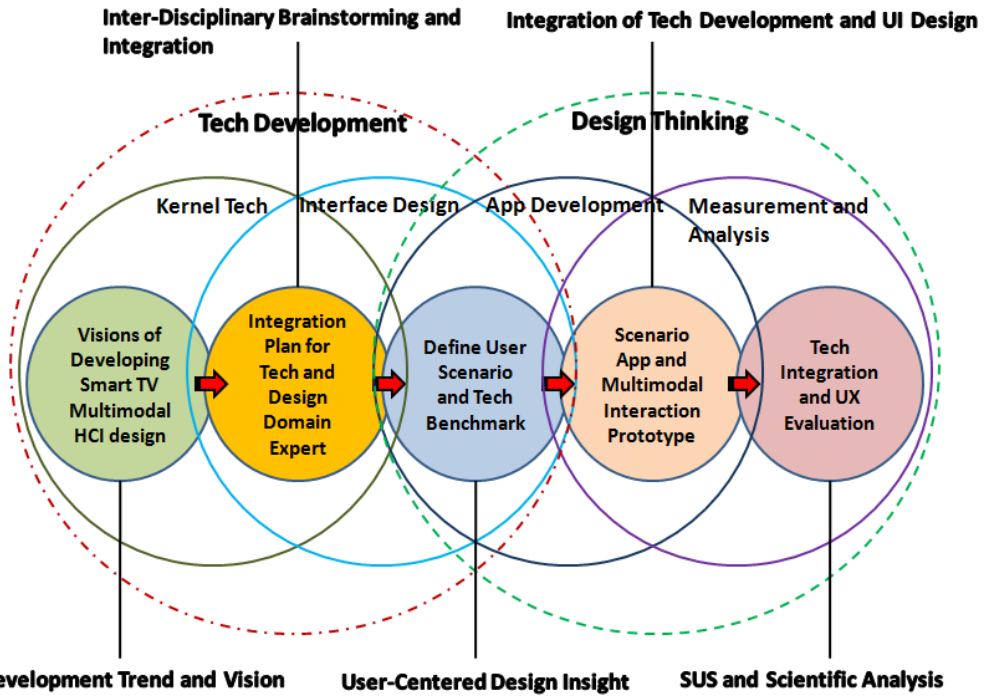


Fig. 1. the Comprehensive Structure of Inter-disciplinary Integration of Design Thinking with Technology Development for Smart TVs

This structure has 5 major phases (Fig. 1). The first phase outlines the vision for developing the Smart TV multimodal HCI design and is based on a review of development trends and visions for Smart TVs. The second phase proposes a plan for inter-disciplinary integration of domain experts in technology and interaction design by holding a service design workshop for brainstorming. The third phase, which is the section for integration, defines the user scenario and technology benchmark with user-centered design insights. The fourth phase develops the applications and multimodal interaction prototype by integrating technologies into the user interface. The final phase evaluates user experiences with the prototype on a system usability scale that can collect scientific data (*e.g.*, eye-tracking system) for objective analysis. Evaluation results are then feed back to the inter-disciplinary team to modify and adjust the prototype. This paper will present the results and evaluation of the first three phases.

For practical implementation, this work follows the implementation process (Fig. 2), the details of which are as follows.

Figure 3 shows inter-disciplinary team discussion and character map of service design workshop. This workshop helped the team gain a clear understanding the features of multimodal HCI design. The many ideas generated were then narrowed down from global thoughts into specific and applicable features that meet user requirements and are applicable to technical features development.

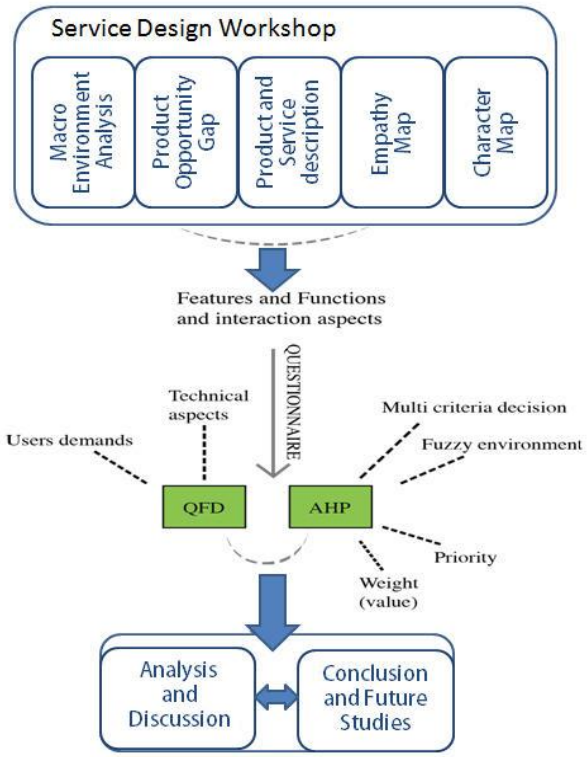


Fig. 2. Implementation Processes

Fig. 3. Inter-disciplinary Team Discussion and Character Map of the Service Design Workshop

A. Quality Function Development (QFD) Matrix

All features derived from the workshop are listed in a QFD matrix. Figure 4 presents the conceptual diagram of the QFD.

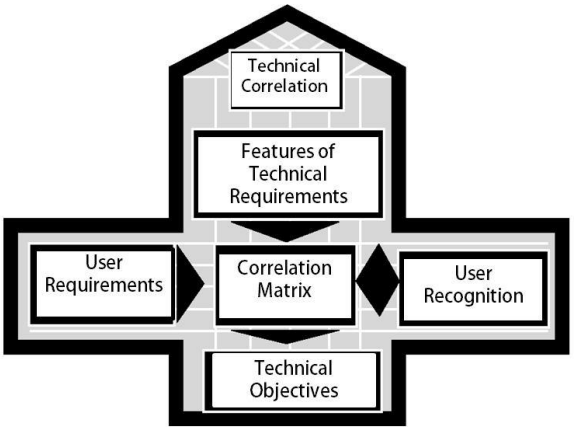


Fig. 4. Conceptual Diagram of Quality Function Deployment

The QFD matrix shows the importance of each feature via correlation analysis of user requirements and features of technical requirements. It also shows user recognition by describing their experiences to competitors by giving a value to their importance. The importance range is 1–5 and their thinking is limited to strong, moderate, or poor. This method tells us how strongly the features (product characteristics) are related to user requirements and reflects the strengths of existing products. This work uses the QFD matrix to systematically list the features of the multimodal HCI design.

B. Analytic Hierarchy Process

The three basic AHP steps in thisresearch are as follows.

1) Describe a complex decision-making problem as a hierarchy.

2) Use pairwise comparison techniques to estimate the relative priority of various elements on each level of the hierarchy.

3) Integrate these priorities to develop an overall evaluation of decision alternatives.

The AHP calculation template provided by Goepel[27] is used for primitive analysis of analytical results. The workbook consists of 20 input worksheets for pairwise comparisons, a sheet for consolidating all assessments, a summary sheet for systematic results, a sheet with reference tables (random index, limits forthe geometric consistency index (GCI), and judgment scales) and a sheet for solving the eigenvalue problem when using the eigenvector method (EVM).

The algorithm and formula used to weight and for pair-wise comparisons are as follows.

a) Multi-criteria decision

In terms of Multi-Criteria Decisions, the AHP uses a three-level hierarchical decision system: the first level considers a decision goal G; on the second level, it has n independent *evaluation criteria*—C_1, C_2,...,C_n, such that $\sum_{i=1}^{n} w(C_i)=1$, where $w(C_i) > 0$, i= 1,2,...,n, $w(C_i)$ is a positive real number—weight, or, relative importance of criterion C_i subject to goal G; on the third level m variants (alternatives) of decision outcomes V_1, V_2,...,V_m are considered, such that again $\sum_{r=1}^{m} w(V_r, C_i)=1$, where $w(V_r, C_i)$ is a non-negative real number—an evaluation (weight) of V_r subject to the criterion C_i, i= 1,2,...,n. This system is characterized by the *super matrix* $\mathbf{W}$, where

$$\mathbf{W}=\begin{bmatrix} \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{W_{21}} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{W_{32}} & \mathbf{I} \end{bmatrix}, \quad (1)$$

where $\mathbf{W_{21}}$ is the $n\times1$ matrix (weighting vector of the criteria), *i.e.*,

$$\mathbf{W}_{21}=\begin{bmatrix} w(C_1) \\ \vdots \\ w(C_n) \end{bmatrix}, \quad (2)$$

and $\mathbf{W_{32}}$ is the $m\times n$ matrix:

$$\mathbf{W}_{32}=\begin{bmatrix} w(C_1,V_1) & \cdots & w(C_n,V_1) \\ \vdots & \cdots & \vdots \\ w(C_1,V_m) & \cdots & w(C_n,V_m) \end{bmatrix}, \quad (3)$$

The variants can be ordered according to these priorities.

$$\mathbf{Z}=\mathbf{W_{32}}\mathbf{W_{21}} \quad (4)$$

In real decision systems with three levels, typical interdependences exist among individual elements of the decision hierarchy (*e.g.*, criteria or variants). Consider now the dependences among the criteria.This system is then given by the super-matrix $\mathbf{W}$:

$$\mathbf{W} = \begin{bmatrix} \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{W_{21}} & \mathbf{W_{22}} & \mathbf{0} \\ \mathbf{0} & \mathbf{W_{32}} & \mathbf{I} \end{bmatrix}, (5)$$

where the interdependences ofthe criteria are characterized by n×n matrix $\mathbf{W_{22}}$:

$$\mathbf{W}_{22} = \begin{bmatrix} w(C_1,C_1) & \cdots & w(C_n,C_1) \\ \vdots & \cdots & \vdots \\ w(C_1,C_n) & \cdots & w(C_n,C_n) \end{bmatrix}.$$

In general, matrix (5) is not column-stochastic; hence, the limiting matrix does not exist. The Stochasticity of this matrix can be retained by additional normalization. A limiting matrix $\mathbf{W_\infty}$ then exists and the vector of weights **Z** can be calculated by formula (6).

$$\mathbf{Z} = \mathbf{W}_{32}\left(\mathbf{I} - \mathbf{W}_{22}\right)^{-1}\mathbf{W}_{21} (6)$$

As matrix $\mathbf{W_{22}}$ is resembles the zero matrix, and the dependences among criteria are generally weak, this result can be replaced by the first four terms of Taylor's expansion:

$$\mathbf{Z} = \mathbf{W}_{32}(\mathbf{I} + \mathbf{W}_{22} + \mathbf{W}_{22}^2 + \mathbf{W}_{22}^3)\mathbf{W}_{21}. \qquad (7)$$

b) Priority Calculation

Priorities $p_i$in each input sheet are calculated using the row geometric mean method (RGMM). With the pairwise *N x N* comparison matrix **A** = a_{ij}

Thus, $r_i = \exp\left[\frac{1}{N}\sum_{j=1}^{N}\ln(a_{ij})\right] = (\prod_{i=1}^{N} a_{ij})^{1/N}$ is calculated

and $p_i = r_i / \sum_{i=1}^{N} r_i$ is normalized.

c) Inconsistencies

To find the most inconsistent comparison, this work looks for the pair*i,j* with

$$\max(\varepsilon_{ij} = a_{ij}\frac{p_j}{p_i})$$

Consistency ratios (CRs) are calculated in all input sheets and in the summary sheet. With $\lambda\, max$, the calculated principal eigenvalue-either based on the priority eigenvector derived by the RGMM in the input sheet or derived by the EVM in the summary sheet-the consistency index (CI) is given as

$$CI = \frac{\lambda_{\max} - N}{N-1}$$

The CR is calculated using $CR = \frac{CI}{RI}$

The Alonson/ Lamata linear fit is applied, yielding *CR*:

$$CR = \frac{\lambda_{\max} - N}{2.7699N - 4.3513 - N}$$

Geometric consistency index (GCI) is calculated using:

$$CGI = \frac{a\sum_{i<j}\ln a_{ij} - \ln\frac{p_i}{p_j}}{(N-1)(N-2)} \qquad (5)$$

d) Aggregation of individual judgments (Consolidation of participants)

The consolidated decision matrix **C** (selected participant "0") combines all ***k*** participants' inputs to obtain the aggregated group result. The weighted geometric mean of the decision matrices elements $a_{ij(k)}$ using the individual decision maker's weight $\boldsymbol{w_k}$, as given in the input sheets, is used:

$$c_{ij} = \exp\frac{\sum_{k=1}^{N} w_k \ln a_{ij(k)}}{\sum_{k=1}^{N} w_k}$$

e) AHP consensus indicator

The AHP consensus is calculated in the summary sheet based on the RGMM results of all inputs using Shannon alpha and beta entropy. The consensus indicator ranges from 0% (no consensus) to 100% (consensus).

AHP consensus indicator S*

$S^* = \left[M - \exp(H_{\alpha\min})/\exp(H_{\gamma\max})\right]/\left[1 - \exp(H_{\alpha\min})/\exp(H_{\gamma\max})\right]$with $M = 1/\exp(H_\beta)$

where? $H_{\alpha,\beta,\gamma}$ is the α, β, γ Shannon entropy for the priorities of all *K* decision makers/participants.

Shannon alpha entropy $H_\alpha = \frac{1}{K}\sum_{j=1}^{K}\sum_{i=1}^{N} - p_{ij}\ln p_{ij}$

Shannon gamma entropy $H_\gamma = \sum_{j=1}^{K} - \overline{p}_j \ln \overline{p}_j$

With $\overline{p}_j = \frac{1}{N}\sum_{i=1}^{N} p_{ij}$

Shannon beta entropy $H_\beta = H_\gamma - H_\alpha$

One must adjust for the maximum score c_{max} of the AHP scale and

$$H_{\alpha\min} = -\frac{c_{\max}}{N + c_{\max} - 1}\ln\left(\frac{c_{\max}}{N + c_{\max} - 1}\right) - (N-1)\frac{1}{N + c_{\max} - 1}\ln\left(\frac{1}{N + c_{\max} - 1}\right)$$

$$H_{\gamma\max} = (N-K)\left(-\frac{1}{c_{\max} + N - 1}\right)\ln\left(\frac{1}{c_{\max} + N - 1}\right) - \left(\frac{K + c_{\max} - 1}{N + c_{\max} - 1}\right)\ln\left(\frac{1}{K}\cdot\frac{K + c_{\max} - 1}{N + c_{\max} - 1}\right)$$

where *N* is number of criteria, and K is the number of decision-makers/participants.

IV. Results and Discussion

A. Quality Function Deployment Matrix Results

Figure 5 shows the QFD matrix results. Based on QFD matrix analysis. the smart interactive user interface and privacy settings are two of the most important features of Smart TVs, followed by gesture and voice control, customization of personal settings, and layout adaptation. These visualized results show that the multimodal interaction design is very important to Smart TVs.

In comparison with technical features, gesture recognition

and facial recognition are highly prized by respondents. Privacy via encryption and decryption, and traditional/single sign-in on account management are also required by customers. Respondents agreed that Apple TVs and Smart TVs have user-friendly interfaces. The privacy feature has already been developed by Apple TV, general Smart TV, and Google TV. The QFD matrix results comprehensivelyshow a significant role to help the development. These results are also evaluated and calculated via the AHP. Each criterion is compared to another, such that the importance weight is derived.

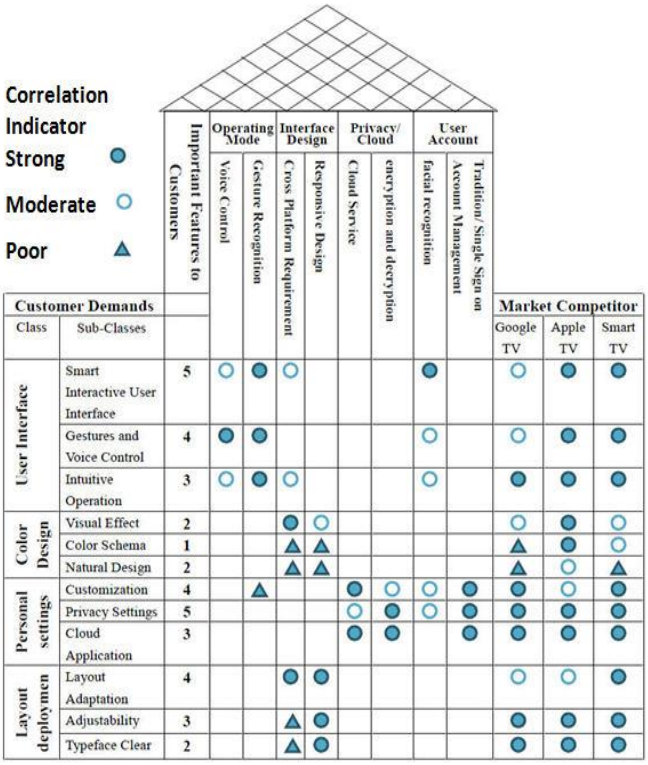


Fig. 5. Quality Function Deployment Matrix Results

B. Analytical Hierarchy Process Results

Features in the QFD matrix are further processed as criteria in a questionnaire. To collect pairwise comparison results, 30 questionnaires were dispatched to inter-disciplinary experts, including faculty, researchers, and professionals in the fields of computer science, electronic engineering, and interaction design.

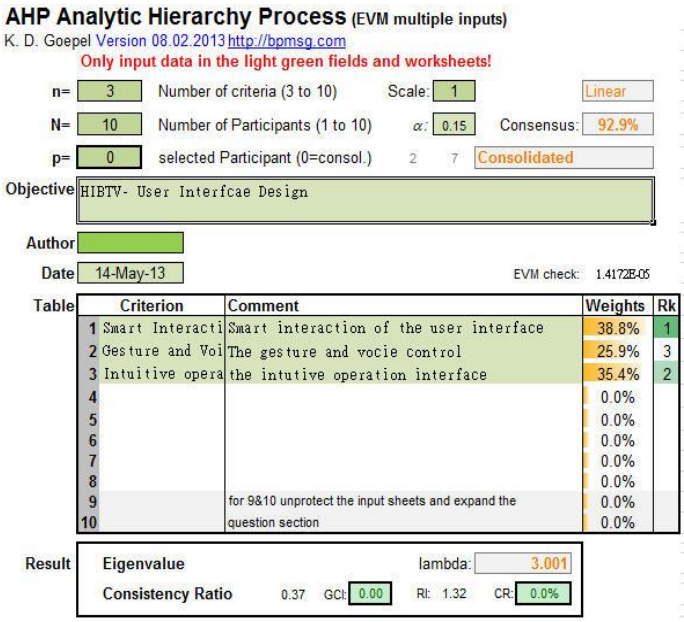


Fig. 6. Summary of the User Interface Class in the User Requirement Category—AHP Results

The quantitative results are then applied in the AHP template to weight and prioritize each feature. Figure 6 shows the table of analytical results for the user interface class in the category of customer demands. According to the ranking, the smart interactive user interface is followed by intuitive operation, and gesture and voice control.

Next, one must turn this matrix into a ranking of criteria (Fig. 7). According to Saaty[21], the eigenvector solution was the best approach. The computed eigenvector gives the relative ranking of criteria. The most important criterion is *smart interactive user interface*(38.8.%), followed by *intuitive operation* (35.4%), and gesture and voice control(25.9%) (Fig. 4). Also, the CR is <1.5%, meaning the ranking is credible.

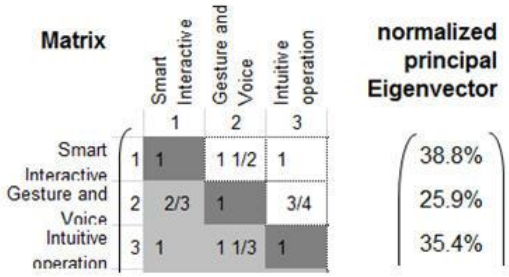


Fig. 7. Normalized Principle Eigenvector Table

The weight scale from the QFD matrix and weights and ranking by the AHP method are correlated (Table 1).

As shown in Table 1, there is a correlation between the QFD Matrix's weight scale and the weights and ranking from AHP method. The most demanded feature of Smart TV HCI is the user interface (31%) then the visual design. (27%). There are not much difference between the personal setting (22%) and layout development (20%). However, the top 5 design priority for Smart TV HCI design feature are: layout adaptation (46%), smart interactive user interface (39%), personal customization setting (39%), nature design in visual design (36%), user interface intuitive operation (35%). The top 3 feature from the AHP are similar to the QFD weight. However, the fourth and fifth design feature priority are very different from OFD results. The gesture and voice control feature, has a high priority in the QFD matrix, differing markedly from its low weight by the AHP method. The likely reason is that this customer requirement differs from the technical perspective. Additional efforts are needed in gesture and voice control when designing multimodal interaction for Smart TVs. The results show some guideline for industry to the development of Smart TV HCI design.

TABLE I. COMPARISON OF QFD AND AHP RESULTS

Class	Smart TV HCI Design Features	QFD Weights	AHP Weights	AHP Overall Ranking
User Interface (31%)	Smart Interactive User Interface	**5**	39%	**2**
	Gestures and Voice Control	**4**	26%	11
	Intuitive operation	3	35%	**5**
Visual Design (27%)	Visual Effect	2	31%	8
	Color Brightness	1	33%	6
	Natural Design	2	36%	**4**
Personal Settings (22%)	Customization	**5**	39%	**2**
	Privacy Settings	**4**	32%	7
	Cloud Application	3	29%	9
Layout Deploy-ment (20%)	Adaptive Layout	**4**	46%	**1**
	Clarity	3	28%	10
	Clear typeface	2	26%	11

V. CONCLUSIONS AND FUTURE WORK

The AHP, based on the hierarchy principle, assumes consecutive decomposition of multiple aims with degree increasing toward lower levels. Hierarchy development conforms with the principles of system approaches toward task analysis and can facilitate the process of creation and formalization of Participatory Technology Development (PTD) priorities. One main advantage of the AHP is the determination of subjective criteria and scores based on pairwise comparisons. Another advantage involves the structural organization of problem components. The AHP provides consistent assessment tools, analyzes alternative sensitivities, uses relatively simple mathematic equations, and allows participation of different specialists or groups.

A strong point of the AHP is the independence of its application from the activity sphere. The AHP results show the service design approach is an efficient way for communication among interdisciplinary team members. The proposed two-stage decision-making processes qualitatively analyze and quantitatively assesses the priority and relevance of features derived from service design process. The technique team can then develop a prototype that demonstrates multimodal interaction with confidence, thereby fulfilling user demands.

Three possibilities directions exist for future study.

1) Include raw prices (retail prices) in the QFD matrix method and AHP. This would be comparative, as this criterion may affect user demands. For example, if the Kinect Sensor price is excessive, and users think it is not as effective as, say, the motion leap sensor embedded in a remote control could be considered as a criterion to be evaluated.

2) In-depth understandings of current and existing demands are essential. Failure probability still exists as the AHP does not work well when evaluating quantitative values; it is much better at creating qualitative values.

3) Implementing these methods is acceptable. For future work could identify Smart TV features. If field report results could be evaluates and joined with questionnaire results, this project would generate relevant and effective content.

ACKNOWLEDGEMENT

This work is supported in part by the National Science Council, Republic of China, Taiwan, under the Contract No. NSC 101-2219-E-027-007 and MOST 103-2221-E-027-062-. Ted Knoy is appreciated for his editorial assistance.

REFERENCES

[1] K. Merkel, "Hybrid broadcast broadband TV, the new way to a comprehensive TV experience," in Electronic Media Technology (CEMT), 2011 14th ITG Conference on, 2011, pp. 1-4.

[2] Z. Lukac, et al., "The experience of implementing a hybrid broadcast broadband television on network enabled tv set," in MIPRO, 2011 Proceedings of the 34th International Convention, 2011, pp. 840-844.

[3] L. Beloqui Yuste, et al., "Effective synchronisation of Hybrid Broadcast and Broadband TV," in Consumer Electronics (ICCE), 2012 IEEE International Conference on, 2012, pp. 160-161.

[4] P. Hamisu, et al., "Accessible UI design and multimodal interaction through hybrid TV platforms: towards a virtual-user centered design framework," in Universal Access in Human-Computer Interaction. Users Diversity, ed: Springer, 2011, pp. 32-41.

[5] C. Jung and V. Hahn, "GUIDE–Adaptive user interfaces for accessible hybrid TV applications," in Second W3C Workshop Web & TV, 2011.

[6] K. Ota, et al., "Extraction of Customers' Potential Requirements Using Service Scenario Planning," in Product-Service Integration for Sustainable Solutions, ed: Springer, 2013, pp. 63-74.

[7] L.-Y. Zhai, et al., "A rough set based QFD approach to the management of imprecise design information in product development," Advanced Engineering Informatics, vol. 23, pp. 222-228, 2009.

[8] S. Desai, et al., "Material and process selection in product design using decision–making technique (AHP)," European Journal of Industrial Engineering, vol. 6, pp. 322-346, 2012.

[9] O. S. Vaidya and S. Kumar, "Analytic hierarchy process: An overview of applications," European Journal of operational research, vol. 169, pp. 1-29, 2006.

[10] S. Moritz, "Service design: practical access to an evolving field," Cologne, Germany: Köln International School of Design, 2005.

[11] B. Moggridge and B. Atkinson, Designing interactions: MIT press Cambridge, 2007.

[12] L. G. Zomerdijk and C. A. Voss, "Service design for experience-centric services," Journal of Service Research, vol. 13, pp. 67-82, 2010.

[13] Z. Obrenovic and D. Starcevic, "Modeling multimodal human-computer interaction," Computer, vol. 37, pp. 65-72, 2004.

[14] S. Oviatt, "Ten myths of multimodal interaction," Communications of the ACM, vol. 42, pp. 74-81, 1999.

[15] J. Blomkvist and S. Holmlid, "Service designers on including stakeholders in service prototyping," presented at the Include 2011: Sixth International conference on Inclusive Design, London, UK, 2011.

[16] J. Blomkvist and S. Holmlid, "Service prototyping according to service design practitioners," presented at the Second Nordic Conference on Service Design and Service Innovation, Linköping, Sweden, 2010.

[17] Y. Akao, "QFD: Past, present, and future," in International Symposium on QFD, 1997, pp. 1-12.

[18] H. Raharjo, et al., "On integrating Kano's model dynamics into QFD for multiple product design," Quality and Reliability Engineering International, vol. 26, pp. 351-363, 2010.

[19] T. L. Saaty and L. G. Vargas, "How to Make a Decision," in Models, Methods, Concepts & Applications of the Analytic Hierarchy Process, ed: Springer, 2012, pp. 1-21.

[20] T. L. Saaty, "Decision-making with the AHP: Why is the principal eigenvector necessary," European Journal of operational research, vol. 145, pp. 85-91, 2003.

[21] T. L. Saaty, "An Exposition on the AHP in Reply to the Paper" Remarks on the Analytic Hierarchy Process"," Management science, vol. 36, pp. 259-268, 1990.

[22] N. Hanumaiah, et al., "Rapid hard tooling process selection using QFD-AHP methodology," Journal of Manufacturing Technology Management, vol. 17, pp. 332-350, 2006.

[23] V. Dubrovin and N. Mironova, "Usage of the Analytic Hierarchy Process for Production Optimization," in Modern Problems of Radio Engineering, Telecommunications, and Computer Science, 2006. TCSET 2006. International Conference, 2006, pp. 576-577.

[24] F. De Felice and A. Petrillo, "A multiple choice decision analysis: an integrated QFD–AHP model for the assessment of customer needs," International Journal of Engineering, Science and Technology, vol. 2, 2011.

[25] P. C. Gupta, et al., "Evaluation and selection methodology for an innovative product design concepts," International Journal of Engineering Science and Technology (IJEST), vol. 3, pp. 3553-3561, 2011.

[26] A. Osterwalder and Y. Pigneur, "Business model generation–a handbook for visionaires, game changers, and challengers," NewYerk Wiley, 2010.

[27] K. Goepel, "Implementing the analytic hierarchy process as a standard method for multi-criteria decision making in corporate enterprises – a new AHP excel template with multiple inputs. ," presented at the The international symposium on the analytic hierarchy process, Kuala Lumpur, 2013.

A General Model for Similarity Measurement between Objects

Manh Hung Nguyen
[1]Posts and Telecommunications Institute of Technology (PTIT)
Hanoi, Vietnam
[2]UMI UMMISCO 209 (IRD/UPMC), Hanoi, Vietnam

Thi Hoi Nguyen
Vietnam Commercial University, Hanoi, Vietnam

***Abstract*—The problem to detect the similarity or the difference between objects are faced regularly in several domains of applications such as e-commerce, social network, expert system, data mining, decision support system, etc. This paper introduces a general model for measuring the similarity between objects based on their attributes. In this model, the similarity on each attribute is defined with different natures and kinds of attributes. This makes our model is general and enables to apply the model in several domains of application. We also present the applying of the model into two applications in social network and e-commerce situations.**

***Keywords*—object similarity; multiple attributes similarity; similarity measurement; decision support.**

I. Introduction

The problem to detect the similarity or the difference between objects are faced regularly in several domains of applications such as e-commerce, social network, expert system, data mining, decision support system, etc. There are many model proposed to measure the similarity between objects in these kinds of applications. For instances, D. Lin [3] proposed a model to measure the similarity between any two objects based on information-theoretic approach. Sayal and Kumar [10] proposed a model on clustering categorical attributes of relational data set types making use of the property of functional dependency as parameter to measure similarity. Reddy and Krishnaiah [9] proposed a similarity measure known as multi-viewpoint based similarity measure to ensure the clusters show all relationships among objects. Honko [1] proposed and investigated several similarity measures on complex structured objects. The objects are understood as examples of a target relation, and they are expressed in a first-order logic language. Meanwhile some other proposed some metrics to measure the similarity between profiles [8], [7]; similarity between objects based on images [5]; similarity between two trajectories [4]; or similarity between texts [6], [2], etc.

This paper introduces a general model for measuring the similarity between objects based on their attributes. In this model, the similarity on each attribute is defined with different kinds of attributes. This makes our model is general and enables to apply the model in several domains of application.

The paper is organised as follows: Section II presents the general similarity model. Section III presents some case studies for the proposed similarity model. Section IV is the conclusion and perspectives.

II. A model for measurement of similarity between objects

Without loss of generality, we assume that there are n concerned features $\{a^1, a^2, ...a^n\}$, which are attributes of considered object p, to measure the similarity between two objects. There are two steps as follows:

- Step 1: estimate the similarity on each considered feature and normalised it into the unit interval $[0, 1]$.
- Step 2: the similarity between two agents is then estimated by averaging the similarity between them on all considered features.

A. Similarity on each feature

The similarity on each considered feature of object is differently estimated based on the kind of feature. We distinguish five kind of feature:

- Feature whose value is a single number
- Feature whose value is a single string
- Feature whose value is an interval of number
- Feature whose value is a single matching
- Feature whose value is a set of ordered discrete numbers (a vector)
- Feature whose value is a set of non-ordered discrete numbers
- Feature whose value is a set of strings

Note that in case that the feature value is a kind of object, we could recursively apply this model (with two steps) to estimate the similarity between the two object to have the similarity of the feature.

1) Feature whose value is a single number: For this kind of feature, we define a possible interval, call $[MIN, MAX]$, for the value of the feature. It means that the value of the feature is acceptable if only if it is inside a given interval. Therefore, suppose that a_i^k, a_j^k are two single number values on the features a^k, of two objects i and j, respectively. The similarity between object i and j $(i, j \in A)$ on feature a^k is defined by the formula:

$$s_{ij}^k = 1 - \frac{\mid a_i^k - a_j^k \mid}{MAX - MIN} \quad (1)$$

For example, the feature of the *age* of a *seller* in an e-commerce application has a possible interval value of $[0, 100]$. So, if both seller i and j are 30 years old, then their similarity on the feature *age* is 1.00 (100%); if seller i is 30 years old, and seller j is 40 years old, then their similarity on the feature *age* is 0.90 (90%).

This computation is also applied for the feature whose value is a single date time.

2) Feature whose value is a single string: Suppose that a_i^k, a_j^k are two single string values on the features a^k, of two agents i and j, respectively. Let $length_i^k$, $length_j^k$ are the length of the single string value of the features a_i^k and a_j^k, respectively, counted by words. And $length_{ij}^k$ is the length of the longest sub-string between a_i^k and a_j^k, counted by word. The similarity between agent i and agent j $(i, j \in A)$ on feature a^k is defined by the equation:

$$s_{ij}^k = \frac{2 * length_{ij}^k}{length_i^k + length_j^k}$$

For example, considering the feature *name* of two agents: "*Eton John*" (length = 2 words) and "*John Lennon*" (length = 2 words), the longest sub-string of these two names is "*John*" (length = 1), then their similarity on the feature *name* is $2 * 1/(2 + 2) = 0.500$ (50.0%). Meanwhile, the similarity on the feature *name* of "*Eton John*" (length = 2) and "*John*" (length = 1) is $2 * 1/(2 + 1) = 0.667$ (66.7%).

3) Feature whose value is a single matching: The value of this kind of feature could be a single number, single boolean value, or single string. But the matching is strictly binary: the similarity is 1 when the two values are identical; 0 when they are different.

Suppose that a_i^k, a_j^k are two single matching values on the features a^k, of two objects i and j, respectively. The similarity between object i and object j $(i, j \in A)$ on feature a^k is defined by the formula:

$$s_{ij}^k = \begin{cases} 1 \text{ if } a_i^k = a_j^k \\ 0 \text{ if } a_i^k \neq a_j^k \end{cases} \quad (2)$$

For example, considering the feature *original city* of user X is "*Paris*", the user Y is "*Paris*", and the user Z is "*London*", then similarity on the feature *original city* between X and Y is 1. Meanwhile, the similarity on the same feature between X and Z is 0.

4) Feature whose value is an interval of number: Suppose that $a_i^k = [x_1, x_2]$, $a_j^k = [y_1, y_2]$ are two interval values on the features a^k, of two objects i and j, respectively. And $[z_1, z_2]$ is the intersection interval of $[x_1, x_2]$ and $[y_1, y_2]$. The similarity between object i and object j $(i, j \in A)$ on feature a^k is defined by the formula:

$$s_{ij}^k = \frac{2 * (z_2 - z_1)}{(x_2 - x_1) + (y_2 - y_1)} \quad (3)$$

In the case that the intersection interval of $[x_1, x_2]$ and $[y_1, y_2]$ is empty, then $s_{ij}^k = 0$. This is also applied for the feature whose value is a time duration.

For example, considering the feature *price interval of preference* of a *seller* in an e-commerce application. If seller i prefers the price between $\$100 - \$300$, and seller j prefers that between $\$200 - \$400$, then the intersection interval between them is $\$200 - \$300$, so the similarity on this feature between these two sellers is $2 * (300 - 200)/((300 - 100) + (400 - 200)) = 0.50$ (50%).

5) Feature whose value is a set of ordered discrete numbers (a vector): Suppose that $a_i^k = (x_1, x_2, ...x_n)$, $a_j^k = (y_1, y_2, ...y_n)$ (n is the size of vector) are two vector values on the features a^k, of two objects i and j, respectively. And the value in each dimension of the vector is limited in an acceptable interval $[MIN, MAX]$. The similarity between object i and object j $(i, j \in A)$ on feature a^k is defined by the formula:

$$s_{ij}^k = 1 - \frac{1}{n} \frac{\sum_{v=1}^{n} \mid x_v - y_v \mid}{(MAX - MIN)} \quad (4)$$

For example, the feature *position* of a robot is represented in a 3-dimensions space whose the limit in each dimension is in an interval $[0, 10]$. If the robot i is at the position $(0, 3, 7)$, and robot j at the position $(6, 10, 2)$ then the similarity on the feature *position* between them is 0.40 (40%).

6) Feature whose value is a set of non-ordered discrete numbers: Suppose that $a_i^k = \{x_1, x_2, ...x_n\}$, $a_j^k = \{y_1, y_2, ...y_m\}$ (n,m are the size of set) are two set values on the features a^k, of two objects i and j, respectively. And the value in each element of the set is limited in an acceptable interval $[MIN, MAX]$. The similarity between object i and object j $(i, j \in A)$ on feature a^k is estimated as follow:

- Sort the set a_i^k in increasing order such that $a_i^k = \{x'_1 \leqslant x'_2 \leqslant ... \leqslant x_n\}$
- Sort the set a_j^k in increasing order such that $a_j^k = \{y'_1 \leqslant y'_2 \leqslant ... \leqslant y_m\}$
- Without lost any generalisation, we suppose that $n \leqslant m$, the similarity between object i and object j $(i, j \in A)$ on feature a^k is defined by the formula:

$$s_{ij}^k = 1 - \frac{1}{m} \left(\frac{\sum_{v=1}^{n} \mid x'_v - y'_v \mid}{(MAX - MIN)} + (m - n) \right) \quad (5)$$

7) Feature whose value is a set of strings: Suppose that a_i^k, a_j^k are two values of type of set of strings on the features a^k, of two objects i and j, respectively. Let $size_i^k$, $size_j^k$ are the size of the set value of the features a_i^k and a_j^k, respectively. And $size_{ij}^k$ is the size of the intersection set of a_i^k and a_j^k. The similarity between object i and object j $(i, j \in A)$ on feature a^k is defined by the formula:

$$s_{ij}^k = \frac{2 * size_{ij}^k}{size_i^k + size_j^k} \quad (6)$$

For example, in the same application of e-commerce, the feature *favorite leisure* of seller i is a set of {*play football, travel, shopping*} (size = 3) and that of seller j is {*travel, play*

football, lecture, play tennis} (size = 4), then the intersection set of the two values is {*play football, travel*} (size = 2), so the similarity of i and j on this feature is $2*2/(3+4) = 0.57$ (57%).

It is easy to prove that all possible values of s_{ij}^k are lied in the interval $[0, 1]$. It means that, after this step, all similarities between the two agents on each feature are normalised into the unit interval. This normalisation enables us to avoid the domination of some feature whose value domain is vast on other features whose value domain is tight.

B. Similarity between objects

Once the similarities between two objects on each feature are estimated, the similarity between the two objects is then estimated by a weighted average aggregation the similarity between them on all considered features as follow:

- Sorting the similarities on each feature by the decreasing of the important level of the feature. Without lost of generalisation, the important level of considered features is decreased from feature 1 to feature n, then the similarities on each feature is $\{s_{ij}^1, s_{ij}^2, ...s_{ij}^n\}$.
- Choosing a weight vector $w = (w_1, w_2, ...w_n)$, where w_k is the weight of the k^{th} sorted feature such that:

$$\begin{cases} w_{k_1} \geqslant w_{k_2} \text{ if } k_1 < k_2 \\ \sum_{k=1}^{n} w_k = 1 \end{cases} \quad (7)$$

- The similarity between object i and object j is:

$$s_{ij} = \sum_{k=1}^{n} w^k * s_{ij}^k \quad (8)$$

where w^k, s_{ij}^k are respectively the weight of the features a^k and the similarity on the feature a^k between object i and object j.

The usage of the weighted average operator leads this formula more flexible and generic. And the application designer could choose their own weight vector to customise the formula such that it is suitable for the nature of their application.

The weight vector is decreasing from head to tail. This corresponds to the decreased order of important level of sorted feature. This vector may be computed by means of Regular Decreasing Monotone (RDM) linguistic quantifier (Zadeh [12], Yager [11]) as follows:

The function $Q : [0, 1] \to [0, 1]$ is a Regular Decreasing Monotone linguistic quantifier, denote RDM, if and only if it satisfies the following conditions:

$$(i)\ Q(0) = 1$$
$$(ii)\ Q(1) = 0$$
$$(iii)\ Q(i_1) \geqslant Q(i_2) \text{ if } i_1 < i_2.$$

For example, the following functions are RDM:

$$(a)\ Q(x) = (1 - x)^m \text{ with } m \geq 1$$
$$(b)\ Q(x) = 1 - \sqrt{1 - (1 - x)^2}.$$

Suppose that Q is a RDM function, the vector w could be generated by function Q as follow:

$$w_i = Q\left(\frac{i-1}{n}\right) - Q\left(\frac{i}{n}\right) \text{ for } i = 1, ..., n$$

III. Case studies

In this section, we present the two potential applications of the proposed model: (i) applying the model to detect the similarity among user profiles in social network, and (ii) applying the model to choose the best product in an e-commerce system.

A. Detecting user profile similarity in social network

Nowadays, the rapid growth of social networks raises several related problems such as: how to find an account of an user on a social network that we know some pies of information about him; how to detect two (or more than two) accounts on two different social network are belong to an unique person; how to regroup the users of a social network into a set of separated groups;etc. This class of problems could be solved by applying the similarity model in this paper to estimate the similarity on user profiles of social networks.

For instance, let consider the problem to detect two (or more than two) accounts on two different social network are belong to an unique person. We could measure the similarity of each potential profile to the considered profile. The one having the highest similarity could be considered as the secondary profile of the considered user. In order to estimate the similarity of profiles, we consider a profile for *social network user* with following features, in the decreased order of important level:

- *Name*: The full display name of user. This feature is a kind of single string value.
- *Age*: The age of user. This feature is a kind of single number value.
- *Sex*: The sex of the user. This feature is a kind of single matching value.
- *Leisure of favorite*: The user preference of leisure. This feature is a kind of string set value.
- *Original city*: The original city of user. This feature is a kind of single matching value.
- *Work place*: The name of the company (or school) that the user is working for (or studying in, respectively). This feature is a kind of single matching value.

Assume that we are considering an user with profile attribute values are $p_0 =$ (*"Eton John", 62, male,* {*music, cinema*}*, "London", "Global Music"*). And we need to estimate the similarity of five profiles to that of Eton John as follow:

$p_1 =$ (*"John Eton", 60, male,* {*music, sport*}*, "London", "World Music"*)

$p_2 =$ (*"Eton John", 62, male,* {*cinema, sport*}*, "London", "Global Cinema"*)

$p_3 =$ (*"Eton John", 65, male,* {*sport*}*, "London", "Global Music"*)

TABLE I: Summary of similarity of five profiles compared to the considered profile p_0

Profiles	Name	Age	Sex	Leisure	City	Work place	Similarity
p_0	Eton John	62	male	music, cinema	London	Global Music	
p_1	John Eton	60	male	music, sport	London	World Music	**0.74**
	0.5	0.98	1	0.5	1	0	
p_2	Eton John	62	male	cinema, sport	London	Global Cinema	**0.90**
	1	1	1	0.5	1	0	
p_3	Eton John	65	male	sport	London	Global Music	**0.85**
	1	0.97	1	0	1	1	
p_4	Eton John	55	male	music, cinema	New York	Global Music	**0.90**
	1	0.93	1	1	0	1	
p_5	Eton John	60	male	cinema, sport	London	Global Music	**0.93**
	1	0.98	1	0.5	1	1	

p_4 = (*"Eton John", 55, male,* {*music, cinema*}*, "New York", "Global Music"*)

p_5 = (*"Eton John", 60, male,* {*cinema, sport*}*, "London", "Global Music"*)

The model is applied as follow (Table I):

- Choosing the RDM function $Q(x) = (1-x)^2$ to generate the weight vector of six elements ($n = 6$) corresponding to six considered attributes of profile. Therefore, the values of $Q(0/6)$ to $Q(6/6)$ are: 1, 0.69, 0.44, 0.25, 0.11, 0.03, 0.
- The weight vector is thus: $w = (0.31, 0.25, 0.19, 0.14, 0.08, 0.03)$.
- Comparing the similarities: $s_5 > s_4 \sim s_2 > s_3 > s_1$, so the profile p_5 is considered as the most similar to the considered profile.

B. Choosing the best product in e-commerce

Let consider an e-commerce application of type e-market: there are several sellers in the e-market. Each of them sell several products. Each product has a different set of value on its attributes. There is a buyer who want to buy a product. He has some preference values on each attribute of product. The buyer could contact all sellers in the e-market to ask them to propose some products which satisfy his preference. Assume that each seller proposes at least one product to the buyer. So the buyer receives many potential products of his preference. But he have to choose only one product to buy. The question is which proposed product is the best suitable for the buyer, in regarding his preference?

For this problem, we could apply the proposed model as follow:

- Representing the preference product of the buyer on its preference values on n considered attributes as $p_0 = a_1^0, a_2^0, ...a_n^0$
- Assume that there are N products received from sellers. For each received product $p_i, i = 1..N$, representing it via its attribute values as $p_i = a_1^i, a_2^i, ...a_n^i$
- Estimating the similarity s_{i0} of each received product $p_i, i = 1..N$ with the preference product p_0. This step could be done by applying the proposed model in this paper.
- The received product $k, 1 \leq k \leq N$ which has the highest similarity s_{k0} with the preference product p_0 will be considered as the best suitable product for the given buyer.

For example, the exchanged product in the e-market is laptop, and the buyer consider on a set of eight attributes, in the decreased order of important level, in the personnel point of view of the buyer:

- *price*: The lower the price, the better for buyer. The highest acceptance threshold for this buyer is \$500. The value of this attribute is single number, but the preference value is an interval of [0,500].
- *trademark*: The value of this attribute is single string. The the preference value is a set of strings (*Apple, Dell, Sony*).
- *processor speed*: The higher the processor speed, the better for buyer. The lowest acceptance threshold for this buyer is $3.0GHz$. The value of this attribute is single number, but the preference value is an interval of [3.0,...].
- *RAM capacity*: The higher the RAM capacity, the better for buyer. The lowest acceptance threshold for this buyer is $2.4GB$. The value of this attribute is single number, but the preference value is an interval of [2.4,...].
- *hard disk capacity (HDD)*: The higher the hard disk capacity, the better for buyer. The lowest acceptance threshold for this buyer is $100GB$. The value of this attribute is single number, but the preference value is an interval of [100,...].
- *weight*: The lower the weight, the better for buyer. The highest acceptance threshold for this buyer is $3.5kg$. The value of this attribute is single number, but the preference value is an interval of [0, 3.5].
- *screen size*: The higher the screen size, the better for buyer. The lowest acceptance threshold for this buyer is $15inches$. The value of this attribute is single number, but the preference value is an interval of [15,...].
- *color*: The value of this attribute is single string. The the preference value is a set of strings (*Black, White*).

In summary, the preference values on attributes are p_0 = ($\leq$ *500,* {*Apple, Dell, Sony*}*,* $\geq$ *3.0,* $\geq$ *2.4,* $\geq$ *100,* $\leq$ *3.5,* $\geq$ *15,* {*Black, White*}). Assume that there are five received products as follow:

TABLE II: Summary of similarity of five products compared to the considered product p_0

Products	Price	Trademark	P. speed	RAM	HDD	Weight	S. size	Color	Similarity
p_0	$\leq$ 500	Apple, Dell, Sony	$\geq$ 3.0	$\geq$ 2.4	$\geq$ 100	$\leq$ 3.5	$\geq$ 15	Black, White	
p_1	400	Sony	2.8	2.2	100	2.5	14	Black	**0.65**
	1	1	0	0	1	1	0	1	
p_2	600	Dell	3.0	2.4	100	3.5	14	Black	**0.73**
	0	1	1	1	1	1	0	1	
p_3	700	Apple	3.5	2.8	150	2.0	15	White	**0.77**
	0	1	1	1	1	1	1	1	
p_4	500	Acer	3.0	2.4	100	3.8	15	Red	**0.69**
	1	0	1	1	1	0	1	0	
p_5	500	Apple	2.8	2.4	80	3.5	14	White	**0.68**
	1	1	0	1	0	1	0	1	

$p_1 =$ (*400, Sony, 2.8, 2.2, 100, 2.5, 14, Black*)

$p_2 =$ (*600, Dell, 3.0, 2.4, 100, 3.5, 14, Black*)

$p_3 =$ (*700, Apple, 3.5, 2.8, 150, 2.0, 15, White*)

$p_4 =$ (*500, Acer, 3.0, 2.4, 100, 3.8, 15, Red*)

$p_5 =$ (*500, Apple, 2.8, 2.4, 80, 3.5, 14, White*)

The model is applied as follow (Table.II):

- Choosing the RDM function $Q(x) = (1-x)^2$ to generate the weight vector of eight elements ($n = 8$) corresponding to eight considered attributes of the product. Therefore, the values of $Q(0/8)$ to $Q(8/8)$ are: 1, 0.77, 0.56, 0.39, 0.25, 0.14, 0.06, 0.02, 0.
- The weight vector is thus: $w = (0.23, 0.21, 0.17, 0.14, 0.11, 0.08, 0.04, 0.02)$.
- Comparing the similarities: $s_3 > s_2 > s_4 > s_5 > s_1$, so the product p_3 is considered as the best suitable product to the buyer preference.

IV. Conclusions

In this paper, we present a model for estimating the semantic similarity between two objects based on their attributes or features via two steps. Firstly, the model estimates the similarity, between the two objects, on each feature and the results are normalised into the interval [0,1]. Secondly, the similarity between the two object is estimated by a weighted aggregation from the similarities on all considered features. This model could be applied into several applications to help some member of the system to choose the best suitable object from a set of potential objects considered such as find the most closed user profile in social network, choose the best product in an e-commerce application.

In the near future we will extend this model to compare the similarity between the behavior of users.

References

[1] Piotr Honko. Description and classification of complex structured objects by applying similarity measures. *International Journal of Approximate Reasoning*, 49(3):539–554, 2008.

[2] Glen Jeh and Jennifer Widom. Simrank: A measure of structural-context similarity. In *Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '02, pages 538–543, New York, NY, USA, 2002. ACM.

[3] Dekang Lin. An information-theoretic definition of similarity. In *Proc. 15th International Conf. on Machine Learning*, pages 296–304. Morgan Kaufmann, San Francisco, CA, 1998.

[4] Hechen Liu and Markus Schneider. Similarity measurement of moving object trajectories. In *Proceedings of the Third ACM SIGSPATIAL International Workshop on GeoStreaming*, IWGS '12, pages 19–22, New York, NY, USA, 2012. ACM.

[5] E. Nowak and F. Jurie. Learning Visual Similarity Measures for Comparing Never Seen Objects. *Computer Vision and Pattern Recognition, 2007. CVPR '07. IEEE Conference on*, pages 1–8, June 2007.

[6] Massimo Paolucci, Takahiro Kawamura, Terry R. Payne, and Katia P. Sycara. Semantic matching of web services capabilities. In *Proceedings of the First International Semantic Web Conference on The Semantic Web*, ISWC '02, pages 333–347, London, UK, 2002. Springer-Verlag.

[7] Olga Peled, Michael Fire, Lior Rokach, and Yuval Elovici. Entity matching in online social networks. *Social Computing / IEEE International Conference on Privacy, Security, Risk and Trust, 2010 IEEE International Conference on*, 0:339–344, 2013.

[8] Elie Raad, Richard Chbeir, and Albert Dipanda. User profile matching in social networks. In *Proceedings of the 2010 13th International Conference on Network-Based Information Systems*, NBIS '10, pages 297–304, Washington, DC, USA, 2010. IEEE Computer Society.

[9] Gaddam Saidi Reddy and Dr.R.V.Krishnaiah. A novel similarity measure for clustering categorical data sets. *IOSR Journal of Computer Engineering (IOSRJCE)*, 4(6):37–42, 2012.

[10] Rishi Sayal and V. Vijay Kumar. A novel similarity measure for clustering categorical data sets. *International Journal of Computer Applications*, 17(1):25–30, March 2011. Published by Foundation of Computer Science.

[11] R. R. Yager. Nonmonotonic OWA operators. *Soft Computing - A Fusion of Foundations, Methodologies and Applications*, 3(3):187–196, 1999.

[12] L. A. Zadeh. A computational approach to fuzzy quantifiers in natural languages. pages 149–184, 1983.

Analysis of Significant Factors for Dengue Infection Prognosis Using the Random Forest Classifier

A.Shameem Fathima
Research Scholar,
Department of Computer Science and Engineering
Manonmanium Sundaranar University
Tamilnadu , India

D.Manimeglai
Head of the Department,
Department of Information Technology
National Engineering College
Tamilnadu , India

***Abstract*—Random forests have emerged as a versatile and highly accurate classification and regression methodology, requiring little tuning and providing interpretable outputs. Here, we briefly explore the possibility of applying this ensemble supervised machine learning technique to predict the vulnerability for complex disease - Dengue which is often baffled with chikungunya viral fever. This study presents a new-fangled approach to determine the significant prognosis factors in dengue patients. Random forests is used to visualize and determine the significant factors that can differentiate between the dengue patients and the healthy subjects and for constructing a dengue disease survivability prediction model during the boosting process to improve accuracy and stability and to reduce over fitting problems. The presented methodology may be incorporated in a variety of applications such as risk management, tailored health communication and decision support systems in healthcare**

Keywords—Data Mining; Dengue Virus; Machine learning; Random Forest

I. INTRODUCTION

Dengue is a rigorous fever spread by the nibble of an infected mosquito Aedes aegypti . Chikungunya is a crippling viral disease transmitted to humans by infected mosquitoes [1]. It is also an arbovirus that shares the same vector with dengue virus. The disease shares some clinical signs with dengue, and can be misdiagnosed in areas where dengue is common. Thus, in dengue-endemic region, chikungunya is also a significant cause of viral fever causing outbreaks associated with severe morbidity. As these reemerging tropical viral diseases have been increasing in the past several years, several research studies have contributed to investigate factors in diseases [2]. The vital aspects of clinical informatics and public health informatics may be essential to improve the ability to bring basic research findings and evaluate the efficiency of interventions across communities which continues to be beyond the reach of scientists and health professionals.

Presently, highly developed techniques in the fields of data mining, a new stream of methodologies, have come into reality; they provide processes for discovering useful patterns or models from large datasets [3]. One of the most common widely used techniques in data mining is classification. It is used to extract models describing important data classes and to predict the outcome in unseen data at the single point of time [4]. Therefore, in order to aid medical practitioners, predict the accurate outcomes, data mining is needed to process voluminous data available from previously solved cases and to imply the possible treatments based on analyzing the abnormal values of some significant attributes.

Generally intelligent techniques used in dengue fever analysis are fuzzy theory [5], decision trees [6], and Bayesian classifier [7]. Recently Random Forest technique has happened to be an attractive ensemble method in machine learning. As a result, several research studies have successfully applied the algorithm to solve classification problems in object detection, including face recognition, video sequences and signal processing systems [8]. The dataset is collected from various laboratories and hospitals in Tamil Nadu. The main contribution is to provide some experimental insights about the behavior of the variable importance index based on random forests. The performance of the random forests is investigated to generate better perfection models in Dengue survivability. The 10-crossfold validation method, confusion matrix, accuracy, sensitivity, specificity and ROC curve are used to evaluate the dengue virus survivability prediction models.

The remainder of this paper is organized as follows section II introduces the basic concepts of Random Forest. Section III presents the methodologies and experimental design used in this paper. Experiment results and discussions are presented in section IV. The conclusion and outline of future work are given in section V.

II. BASIC CONCEPTS OF RANDOM FOREST

Random Forests [RF] is essentially a data mining package based fundamentally on regression tree analysis [9]. RF tries to perform regression on the specified variables to provide the suitable model. RF uses bootstrapping to produce random trees and it has its own cross validation techniques to validate the model for prediction / classification. Being one of the ensemble learning techniques, Random Forest has been proven to be especially accepted and dominant techniques in the pattern recognition and machine learning for high-dimensional classification [10] and skewed problems [9]. These studies used RF to construct a collection of individual decision tree classifiers which utilized the classification and regression trees (CART) algorithms [11]. The RF model building procedure is essentially the same as a normal classification tree, but with randomness introduced. The procedure is as follows:

1) For the whole set of training data points (predictors and their corresponding response), RF.

2) Each tree when terminal nodes are reached is saved and RF repeats the process. The user specifies how many times this process is repeated (how many trees to grow).

Once the total number of trees is grown the model (or forest) can be saved for subsequent loading in R. RF also supplies the variable importance for each of the predictors in the training data.

Not only is there often a large number of records in the database, but there can also be a large number of fields (attributes, variables); so, the dimensionality of the problem is high. A high-dimensional data set creates problems in terms of increasing the size of the search space for model induction in a combinatorial explosive manner. In addition, it increases the chances that a data mining algorithm will find spurious patterns that are not valid in general. Approaches to this problem include methods to reduce the effective dimensionality of the problem and the use of prior knowledge to identify irrelevant variables. As a classifier, random forest performs an implicit feature selection, using a small subset of "strong variables" for the classification only [12], leading to its superior performance on high dimensional data. The outcome of this implicit feature selection of the random forest can be visualized by the "Gini importance" [9], and can be used as a general indicator of feature relevance. This feature importance score provides a relative ranking of the features, and is – technically – a by-product in the training of the random forest classifier

III. Methodologies and Experiment Design

This paper, focusing on random forests, the increasingly used statistical method for classification and regression problems introduced by Leo Breiman in 2001, proposes to investigate two classical issues of variable selection. The first one is to find important variables for interpretation and the second one is more restrictive and tries to design a good cost-conscious prediction model. In this section, the viral data preparation used in this experiment is first described. Then the performance evaluation methods including accuracy, sensitivity, specificity and Receiver Operating Characteristic (ROC) curve is presented.

A. Dataset

The Dengue survivability data and viral particles in samples of patients clinically suspected for having dengue fever were obtained from several hospitals, King Institute of preventive Medicine and laboratory diagnostic centers in Tamil Nadu, India. The data includes patient information that was diagnosed with dengue during the year 2009-2011. Clinical presentation was recorded from the patients at different stages those during included in the study. The arboviral survivability data consist of nearly 5000 instances and 29 attributes. These variables are widely used in our hospitals for the diagnosis and monitoring of dengue patients .The whole dataset if divided into two classes, 'Dengue positive' in which the patients are suspected for having dengue fever and also on real time PCR result proves to be Dengue positive and Dengue negative – class in which the patients are suspected for having dengue fever but on real time PCR result proves to be negative. All this raw data does not necessarily equates to having useful information; on the contrary, it could lead to an information overflow rather than insight. What doctors need is high-quality support for making decisions. Data mining techniques can be used to extract useful knowledge from clinical data, to provide evidence for and thus support medical decision making Symptoms for chikungunya and dengue are almost identical - high fever, headache, eye ache, joint pain, rashes and lethargy. These viral diseases are characterized by an abrupt onset of fever frequently accompanied by joint pain. Other common signs and symptoms include muscle pain, headache, nausea, fatigue and rash. The joint pain is often very debilitating, but usually lasts for a few days or may be prolonged to weeks. Symptoms appear between 4 and 7 days after the patient has been bitten by the infected mosquito and these include:

- High fever (40°C/ 104°F)
- Joint pain (lower back, ankle, knees, wrists or phalanges)
- Joint swelling
- Rash
- Headache
- Muscle pain
- Nausea
- Fatigue

B. Evaluation methods

For the success of any data mining project, the data and especially the number of attributes play an important role. The more attributes are used, the higher the probability becomes that strong predictors are identified, and non-linearity and multivariate relationship can occur that intelligent techniques can exploit. If number of attributes increases, the density of the data set in pattern space drops exponentially and complexity of models can grow linearly or worse [13]. Complex models (i.e. a large number of parameters) have a higher chance of over fitting to the training data and will not perform well on new data (low generalization), so attribute selection is important. In this experiment, evaluation methods including basic performance measures and ROC curve are applied.

These evaluation methods are based on the confusion matrix. The confusion matrix is a visualization tool commonly used to present performances of classifiers in classification tasks [3]. It is used to show the associations between real class attributes and that of predicted classes. The intensity of effectiveness of the classification model is calculated with the number of correct and incorrect classifications in each possible value of the variables being classified in the confusion matrix [14] (see Fig. 1).

		Predicted Class	
		Dengue Positive	Dengue Negative
Outcome	Dengue Positive	TP	FN
	Dengue Negative	FP	TN

Fig. 1. The Confusion Matrix

The confusion matrix is used to compute true positives (TP), false positives (FP), true negatives (TN) and false negatives (FN), as represented in Fig. 1.

C. Performance Measures

There are three commonly used performance measurements including accuracy, sensitivity and specificity [3]. The accuracy of classifiers is the percentage of correctness of outcome among the test sets exploited in this study as defined in (1). The sensitivity is referred as the true positive rate, and the specificity as the true negative rate. Both sensitivity and specificity used for measuring the factors that affect the performance are presented in (2) and (3), respectively.

$$\text{Accuracy} = \frac{TP+TN}{(TP+FP+TN+FN)} \quad \ldots\ldots\ldots\ldots \quad (1)$$

$$\text{Sensitivity} = \frac{TP}{(TP+FP)} \quad \ldots\ldots\ldots\ldots \quad (2)$$

$$\text{Specificity} = \frac{TN}{(TN+FN)} \quad \ldots\ldots\ldots\ldots \quad (3)$$

The risk rate of the corresponding integrated risk factor associated with each prediction method is reported. It is computed as the ratio of the probability of developing disease among those predicted susceptible to the probability of developing disease among those predicted non-susceptible.

D. Receiver Operating Characteristic (ROC) curve

The Receiver Operating Characteristic curve graphically interprets the performance of the algorithm implemented. It is used as an evaluation criterion for the predictive performance of the classification or the data mining algorithms [15]. ROC curve is a two-dimension graph in which the true positive rate (TPR) (4) is plotted on the Y axis and the false positive rate (FPR) (5) is plotted on the X axis. *TPR* is the true positive value which is the number of correct predictions. *FPR* is the false positive value which is the number of incorrect predictions.

$$TPR = \frac{TP}{TP+FN} \quad \ldots\ldots\ldots\ldots \quad (4)$$

$$FPR = \frac{FP}{TN+FP} \quad \ldots\ldots\ldots\ldots \quad (5)$$

ROC analysis offers more robust evaluation of the relative prediction performance of the models than the tradition comparison of relative error, such as error rate [16].

IV. Results and Discussion

All analyses were carried out using R - a free, cross-platform, open-source statistical analysis language and program. It is also an alternative to expensive commercial statistics software such as SPSS. Packages extend the functionality of R by enabling additional visual capabilities, statistical methods, and discipline-specific functions [17]. The recommended R distribution includes a number of packages in its library. These are collections of functions and data [18]. The base package, the stats package, the datasets package and several other packages, are automatically attached at the beginning of a session**.** Both of the random Forest package, ROCR package, party package and rpart package [17] [18] is frequently used.

For biological research applications, interpretability of results is a key factor in selecting a particular machine learning method. For the experiment results, we are interested in the percentage of correctly classified instances of the algorithm (accuracy percentage) and the number of rules or size of trees produced by the classifiers. For the experimental setup, all the original datasets are entered in to excel sheet and saved as csv file format and imported as input to the R software for analysis. Next, the identified classification technique is implemented and tested on the viral dataset.One part of the data is used to create the classifier, the other part is held out to test the performance of the model on cases that have not been used for training. A more sophisticated internal validation method is cross validation. This procedure will result in a more accurate estimate of the model performance.

For RF analysis, RF classification tree methods (number of trees =500; number of variables tried at each split =5) is used. To measure the importance of predictor variables, the mean decrease in accuracy and Gini index at each node were used. Fig. 2 illustrates the 29 most important variables of each measure. Mean Decrease in Accuracy exploits the margin, defined as the average of (% of votes for true class in the untouched OOB data) - (% of votes for the correct class in the variable-permuted OOB data) over all trees. In other words, the larger the size of the margin, the more important the predictor is. Gini importance is calculated for each variable using the Gini impurity criterion of the resulting subsets of the data at each decision node where the variable was used. Gini impurity is based on the squared probabilities of cases and controls in the two resultant subsets after a split is made using a variable. By definition, the impurity in the resulting subsets must be less than in the parent subset. The Gini index for a given variable is the sum over all trees of the decrease in Gini impurity after each split that involved that variable.

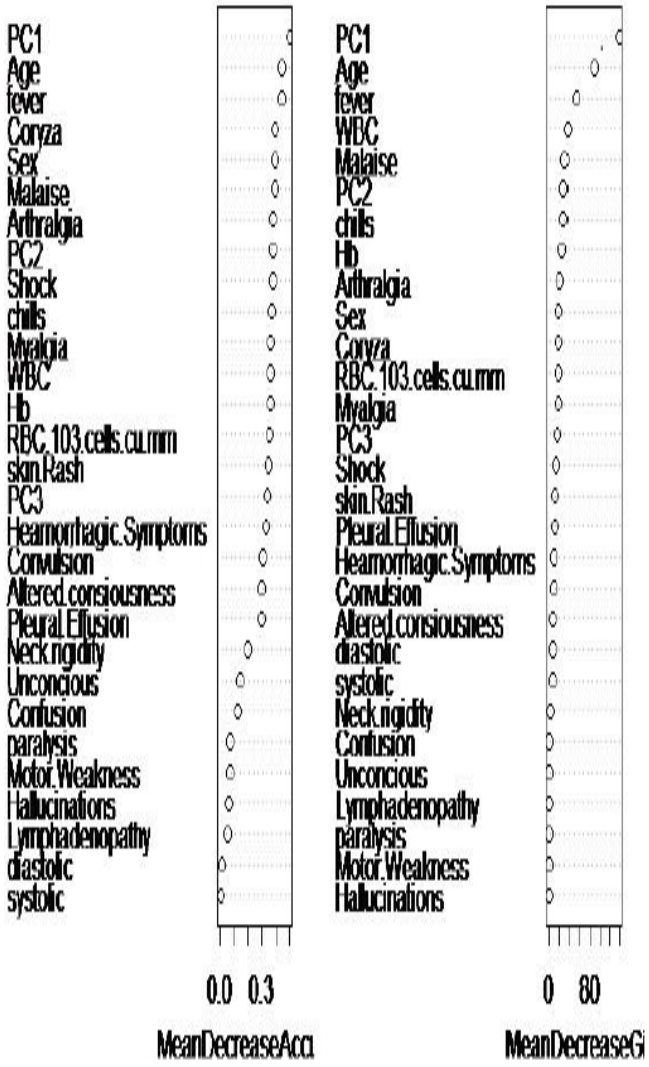


Fig. 2. Variable Importance Plots with Top 29 Variables caption

Left panel contains the 29 most important variables for predicting case control status descending by Mean Decrease Accuracy (average of (% of votes for true class in the untouched OOB data) - (% of votes for the correct class in the variable-permuted OOB data) over all trees). Right panel contains the 29 most important variables descending by Mean Decrease Gini Index (adding up the Gini decrease for each individual variable over all trees). Both the accuracy measure and the Gini index detected the variables which had significant p-values less than 0.0001 for the Fisher's exact test within the 29 most important variables.

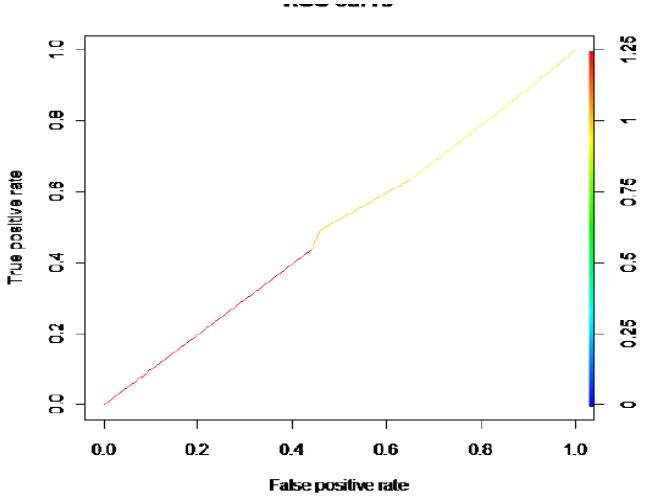


Fig. 3. ROC Curve

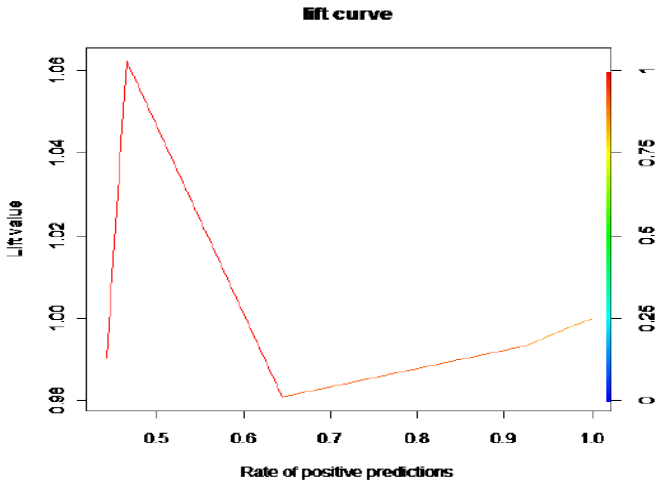


Fig. 4. LIFT CURVE

A predictive model is created using cforest (Breiman's random forests) from the package *party*, to evaluate the predictive model on a separate set of data, and then the performance using ROC curves and a lift chart is plotted. These charts are useful for evaluating model performance in data mining and machine learning. The performance of the model applied to the evaluation set is plotted as an ROC curve and lift chart as seen in Fig 3 and 4 respectively.

Permutation importance, on the other hand, is a reliable measure of variable importance for uncorrelated predictors when sub-sampling without replacement — instead of bootstrap sampling — and unbiased trees are used in the construction of the forest [19]. To meet this aim, conditional permutation is performed in which the importance measure is able to reveal the fake correlation between the response variable and other predictor variables. The results of conditional permutation scheme are shown in Fig.5.

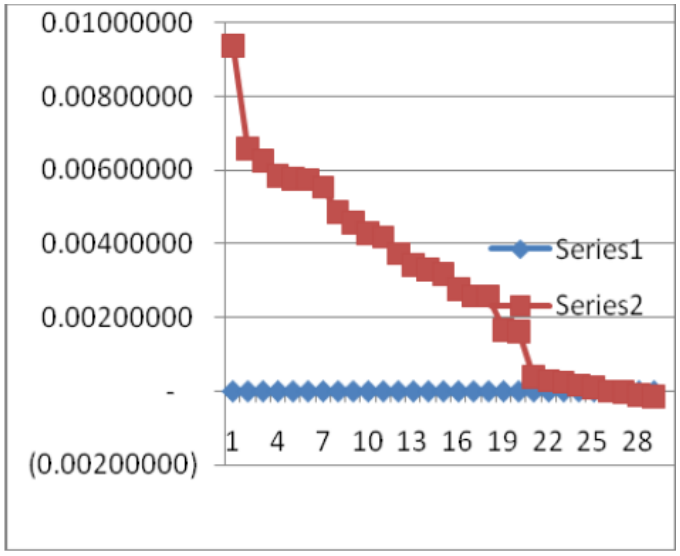


Fig. 5. Plot showing the conditional Importance of each variable

By inferring from these results the most important attributes are identified in the order of Platelet count 1, Malaise, Coryza, Myalgia, Platelet count, Chills, Arthralagia, White blood cells count , Fever .These results are compared with the instantaneous study of the viral diseases by the doctors and virologists reported by the World Health Organization .The report proves that the Patients with dengue had significantly lower platelet, white blood cell (WBC) and Signs of rash and indicators of liver damage, in combination with other variables such as age, myalgia , WBC count, and platelet counts [20]. The findings of this study suggest that several clinical and laboratory measures could potentially distinguish patients with dengue from those with other viral disease. Low platelet count and decreases in WBC and neutrophils were independently associated with the presence of dengue [21] [22] [23].The performance measures obtained by the implemented technique is tabulated and shown below in table 1

TABLE I. PERFORMANCE OF THE SINGLE CLASSIFIER ON THE DATA

RF- Performance measures	
Sensitivity	.9404
Specificity	0.9219
Accuracy	0.9234
Risk Rate	0.519
TPR	0.51
FPR	0.99

As shown from the results, using RF as a base learning algorithm ability of prediction is reduced and the present study highlighted important clinical observations of dengue viruses, to rule out the present confusion and may help to establish a diagnostic algorithm to distinguish dengue from other viral patients. The study also guides in early detection of the viral diseases so that appropriate management may be undertaken to reduce the long-lasting consequences in health. Random forest runtimes are quite fast, and they are able to deal with unbalanced and missing data.

V. CONCLUSION AND FUTURE WORK

Identification of the influential clinical symptoms and laboratory features that help in the diagnosis of dengue fever (DF) in early phase of the illness would aid in designing effective public health management and virological surveillance strategies. Keeping this as our main objective, we develop in this paper a new computational intelligence-based methodology that predicts the diagnosis in real time, minimizing the number of false positives and false negatives.Given its performance, random forest and variable selection using random forest should probably become part of the "standard tool-box" of methods for the analysis of dengue data. The proposed method can be used for variable selection fulfilling the objectives above. Screen plots can be used to recover the important variables that are related to the diagnosis of Dengue, with-out being adversely affected by collinear ties; the proposed method is capable of extracting patterns, but with-out the cooperation and feedback from the medical practitioner, these results would be useless. Besides, this method is not aimed at replacing the medical practitioner and researchers, but rather to complement their invaluable efforts to save more human lives. As for further work, the plan is to investigate the diversity of the number of classifiers such as linear discriminant analysis, logistic regression and support vector machines in this aspect. Another possibility to investigate is using the RF algorithm in larger data sets with scores of attributes. Finally, a comparison of the classifiers ensemble would be of interest.

ACKNOWLEDGMENT

Thanks to the Virology department staff at King Institute of Preventive Medicine, Chennai, Tamilnadu, doctors and microbiologists who provided us with a cosmic amount of viral data needed for our research study and validated our results.

REFERENCES

[1] Chakkaravarthy, V.M., S. Vincent and T. Ambrose,1011. Novel Ap proach of Geographic Information Systems on Recent outbreaks of Chikungunya in Tamil Nadu, India. J. Env.Sci. Tech.4(4):387-394 *(references)*

[2] T. Srinivasan, A. Chandrasekhar, J. Seshadri and J. B. S. Jonathan,―Knowledge discovery in clinical databases with neural network evidence combination,‖ in Proc. International Conference on Intelligent Sensing and Information, 2005, pp. 512-517.

[3] J. Han and M. Kamber, Data mining: concepts and techniques. 2nd.ed. San Francisco: Morgan Kaufmann, Elsevier Science, 2006.

[4] M. T. Skevofilakas, K. S. Nikita, P. H. Templaleksis, K. N. Bir bas, I.G. Kaklamanos and G. N. Bonatsos, ―A decision support system for breast cancer treatment based on data mining tech-nologies and clinical practice guidelines,‖ in IEEE-EMBS the Twenty-SeventhAnnual International Conference on Medicine and Biology Society,2005, pp. 2429-2432.

[5] Parido, A., and P. Bonelli. A new approach to fuzzy classifier systems.In Proceedings of the Fifth International Conference on GeneticAlgorithms. pp. 223–230. 1993.

[6] Hassanien, A.E. Classification and feature selection of breast cancer data based on decision tree algorithm. International Journal of Studies in Informatics and Control Journal, 12(1), 33– 39.2003.

[7] Cheeseman, P., and J. Stutz. Bayesian classification (AutoClass): Theory and results. In U.M. Fayyad, G. Piatetsky–Shapiro P. Smyth and R.Uthurusamy (Eds.), Advances in Knowledge Discovery and DataMining. AAAI Press/MIT Press.1996

[8] M. Zhou and H. Wei, ―Face Verification Using GaborWavelets And AdaBoost,‖ in the Eighteenth International Conference on PatternRecognition, Hong Kong, 2006, pp. 404-407.

[9] L. Breiman, ―Random Forests,‖ J. Machine Learning vol. 45, pp. 5– 32, 2001.

[10] N. Meinshausen, ― Quantile Regression Forests,‖ J. Machine Learning Research, vol. 7, pp. 983–999, 2006.

[11] L. Breiman, J. Friedman, R. Olshen and C. Stone, Classification and regression trees. Wadsworth: Belmont, 1984

[12] Breiman L. Technical Report 670. Technical report, Department of Statistics, University of California, Berkeley, USA; 2004. Consistency for a simple model of random forests.

[13] Bishop, C.M. (1995), Neural Networks for Pattern Recognition, Oxford University Press, Oxford, UK.

[14] P. Cabena, P. Hadjinian, R. Stadler, J. Verhees and A. Zana si, Discovering data mining from concept to implementation. Upper Saddle River, N.J.: Prentice Hall, 1998.

[15] S. Agarwal, T. Graepel, R. Herbrich, S. Har-Peled and D. Roth,―Generalization bounds for the area under the ROC curve,‖ J. Machine Learning Research, vol. 6, pp. 393-425, 2005.

[16] R. O. Duda, D. G. Stork and P. E. Hart, Pattern classification. 2nd ed. New York: Wiley, 2001.

[17] Maindonald, J. H., 2001. Using r for data analysis and graphics, http://wwwmaths.anu.edu.au/~johnm/r/usingR.pdf.

[18] R Core Development Team. *An Introduction to* R.http://cran.r-project.org

[19] C. Strobl, A.-L. Boulesteix, A. Zeileis, and T. Hothorn.Bias in random forest variable importance measures:Illustrations,sources and a solution. BMC Bioinformatics, 8:25, 2007b.

[20] Ageep AK, Malik AA, Elkarsani MS. Clinical presentations and laboratory findings in suspected cases of dengue virus. *Saudi Med J.* 2006;27:1711–1713

[21] Dengue hemorrhagic fever: diagnosis, treatment, prevention and control. 2nd edition. Geneva: World Health Organization 1997.

[22] A. Hapfelmeier and K. Ulm –A new Variable selection approach Using RamdomForests Computational Statistics & Data Analysis, 2013, vol. 60, issue C, pages 50-69

[23] Ranjit S, Kissoon N, Gandhi D, Dayal A, Rajeshwari N, Kamath SR. Early differentiation between dengue and septic shock by comparison of admission hemodynamic, clinical, and laboratory variables: a pilot study. *Pediatr Emerg Care*. 2007;23:368– 375. [PubMed]

[24] (2013). UCI Machine Learning Repository. Available:http://archive.ics.uci.edu/ml/

[25] K. Srinivas, B. K. Rani, and A. Govrdhan, "Applications of data mining techniques in healthcare and prediction of heart attacks," International Journal on Computer Science and Engineering (IJCSE), vol. 2, pp. 250-255, 2010

Confinement for Active Objects

Florian Kammüller
Middlesex University, London UK and
Technische Universit¨at Berlin,
Germany

***Abstract*—In this paper, we provide a formal framework for the security of distributed active objects. Active objects communicate asynchronously implementing method calls via futures. We base the formal framework on a security model that uses a semi-lattice to enable multi-lateral security crucial for distributed architectures. We further provide a security type system for the programming model ASPfun of functional active objects. Type safety and a confinement property are presented. ASPfun thus realizes secure down calls.**

***Keywords*—Distributed active objects, formalization, security type systems**

I. Introduction

Formal models for actor systems become increasingly important for the security analysis of distributed applications. For example, models of organisational structures together with actors provide a basis for the analysis of insider threats, [31], [32].

Active objects define a programming model similar to actors [2] but closely related to object-orientation. An object is an *active object* if it serves as an access point to its own methods and associated (passive) objects and their threads. Consequently, every call to those methods will be from outside. These remote calls are collected in a list of requests. The unit comprising the object's methods and attributes and its current requests is called *activity*. The activity serves as a unit of distribution since it has a data space separate from its environment and can process requests independently. To enable asynchronous communication between distributed active objects, the concept of *futures* – promises for method call values – is used. Active objects are practically implemented in the Java API ProActive [6] developed by Inria and commercialized by its spin-off ActiveEON. Active objects are also a tangible abstraction for distributed information systems beyond just one specific language. ASP [7] is a calculus for active objects. ASP has been simplified into ASP_{fun} – a calculus of *functional* active objects. ASP_{fun} is formalized in Isabelle/HOL [18] thus providing a general automated framework for the exploration of properties of active objects.

In this paper, we use this framework to support security specification and analysis of active objects. The contributions of this paper are (a) the formalization of a novel security model for distributed active objects that supports multi-lateral security, (b) a type system for the static security analysis for ASP_{fun} configurations, (c) preservation and the simple security property of confinement for well-typed configurations, (d) and an argument that secure down calls are possible for ASP_{fun}.

The novel security model [21] is tailored to active objects as it supports decentralized privacy specification of data in distributed entities. This is commonly known as multi-lateral security. To achieve it we break away from the classical dogma of lattices of security classes and use instead semi-lattices. In our model, we implement *confinement*. Every object can remotely access only public (L) methods of other activities. Methods can be specified as private (H) in an activity forbidding direct access. All other methods of objects are assumed to be L, partitioning methods locally into L and H. The security policy further forbids local information flow from H to L. To access an L-method remotely, the containing activity must also be visible to the calling activity in a configuration. In ASP_{fun}, this visibility relation is implemented by activity references. In other active object programming languages, visibility could be given alternatively by an import relation or a registry.

In this paper, we provide an implementation of this security model in the ASP_{fun} framework to illustrate its feasibility and the applicability of the ASP_{fun} framework.

We design a security type system for ASP_{fun} that implements a type check for a security specification of active objects and visibility. We prove the preservation property for type safety of the type system guaranteeing that types are not changed by the evaluation of an ASP_{fun} configuration. The specification of parts of an active object as confined, or private (or H), is possible at the discretion of the user. This specification is entered as a security assignment into the type system; by showing a general theorem that confinement is entailed in well-typedness, we thus know that a well-typed program provides confinement of private methods. Although the confinement property intuitively suffices for security, at this point, a formal security proof is still missing. Moreover, implicit flows may occur. We thus provide a definition of noninterference for active objects. Based on that, we prove that a well-typed configuration does not leak information to active objects below in the hierarchy of the security model, i.e., multi-lateral security holds for well-typed configurations.

Remote method calls in ASP_{fun} have no side-effects. Hence, secure down calls can be made. Confinement provides that no private information is accessed remotely and side-effect freedom guarantees that through the call no information from the caller side is leaked. Side effects are excluded in our formal model ASP_{fun} because it is functional but this can be implemented into the run-time system of other active object languages.

Overview

We first review the semi-lattice for multi-lateral security (Section II-A) and ASP_{fun} (Section II-B) introducing a running example of private sorting (Section II-C). Next, we describe how the semi-lattice model can be applied to active

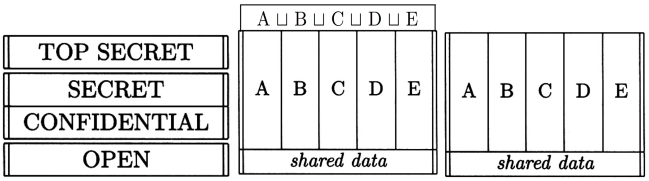


Fig. 1. Joins enable Top control in MLS models

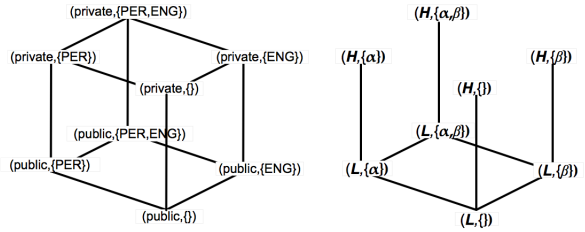


Fig. 2. Taking the top off MLS lattice (left) leads to semi-lattice (right).

objects by instantiating it for ASP_{fun}(Section III). We discuss secure down calls, a distinctive feature of ASP_{fun} enabled by its functional nature and that moreover does not restrict common bi-directional communication patterns. To show the latter point, we present how to implement the Needham-Schroeder Public Key protocol in ASP_{fun}. We describe what we mean by security, i.e., the attacker model and the information flows between active objects through method calls (Section III-C) and illustrate their enforcement on the running example. Following that, we present a type system for the static analysis of a configuration of active objects in ASP_{fun} (Section IV). Properties of this type system are presented (Section V): (a) preservation as a standard result of type safety and (b) confinement. We then define noninterference and multi-lateral security formally to present a soundness theorem, i.e., well-typed configurations are multi-lateral secure. We finish the paper with a related work section and also give some conclusions (Section VI). An Appendix contains sections A ...E with formal details, more examples, and (full) proofs.

II. Prerequisites

A. Semi-Lattice Model for Privacy

We abstract the confinement property known from object oriented languages, e.g., private/public in Java, and use it as a blueprint for a model of privacy in distributed objects. Consider Figure 1: multi-*level* security models support strict hierarchies like military organization (left); multi-*lateral* security [3, Ch. 8] is intended to support a decentralized security world where parties A to E share resources without a strict hierarchy (right) thereby granting privacy at the discretion of each party. But lattice-based security models usually achieve the middle schema: since a lattice has joins, there is a security class A ⊔ B ⊔ C ⊔ D ⊔ E that has unrestricted access to all classes A to E. For a truely decentralized multi-lateral security model this top element is considered harmful. To realize confinement, we exclude the top element by excluding joins from the lattice. We thereby arrive at an algebraic structure called a semi-lattice in which meets always exist but not joins.

Semi-Lattice: The semi-lattice of security classes for active objects is a combination of global and local security lattices. The two lattices are used to classify the methods into groups and objects into hierarchies.

1) Local Classification: The local classification is used to control the information flow inside an object, where methods are called and executed. For every active object there is the public (L) and a private (H) level partitioning the set of this active object's methods. The order relation of the lattice for local classification is the relation $\leq$ defined on $\{L, H\}$ as $\{(L, L), (L, H), (H, H)\}$.

2) Global Classification: The purpose of the global classification is to control the course of information flows between methods of globally distributed objects and lead their information together in a common dominating activity. To remotely access active objects, the key is their identity (we use α, β to denote identities). As classes for the global lattice we use subsets of the set of all activity identities $\mathcal{I}$. These subsets of compartments build the lattice of global classes, the powerset lattice $\mathcal{P}(\mathcal{I})$ over activity identities $\mathcal{I}$.

$$(\mathcal{P}(\mathcal{I}), \cap, \cup, \subseteq, \varnothing, \mathcal{I})$$

In a concrete configuration, the global class label of an activity is the set of activity identities to which access is granted. For example, with respect to the Hasse diagram in Figure 2, an object at global level $\{\alpha, \beta\} \in \mathcal{P}(\mathcal{I})$ can access any part (method) of an object labeled as $\{\beta\}$ or $\{\alpha\}$ or $\{\}$ but only if this part is additionally labeled as L. Vice versa an object at level $\{\alpha\}$ can neither access L nor H parts of objects at level $\{\beta\}$ nor any parts at level $\{\alpha, \beta\}$ but only L parts at level $\{\}$. Thus the classification of parts of an active object needs to combine labels.

3) Combination of Lattices: The security model of the semi-lattice needs to combine the local and global classification scheme. As result, a *security class* is a pair of local and global class (S, δ). We want to impose confinement of methods in order to realize multi-lateral security with our model. Thus, we have to define the combination of the two constituting lattices such that its order relation corresponds to a multi-lateral information flow relation. I.e., private methods of an object are not accessible by any other than the object itself.

Consequently, the new order for security classes is defined as follows. The combined security class ordering for active objects is defined such that a method class (H, δ) dominates (L, δ) and also (L, δ') for all $\delta' \subseteq \delta$ but no other (X, δ_0) dominates (H, δ). The combination of local and global types into pairs gives a partial order

$$CL \equiv (\{L, H\} \times \mathcal{P}(\mathcal{I}), \sqsubseteq)$$

with

$$(S_0, I_0) \sqsubseteq (S_1, I_1) \equiv \begin{pmatrix} S_0 <_S S_1 \vee S_0 = S_1 = L \\ I_0 \subseteq I_1 \end{pmatrix}$$

where the vertical notation $\binom{\phi}{\xi}$ abbreviates $\phi \wedge \xi$ and $<_S = \{(L, H)\}$ denotes the strict ordering on the local security classes. Consequently, meets exist but no joins. The partial order *CL* is thus just a semi-lattice as illustrated by an example in Figure 2 (right).

B. Functional Active Objects: ASP_{fun}

ASP_{fun} uses a slightly extended form of the simplest ς-calculus from the Theory of Objects [1] by distributing ς-calculus objects into activities. The calculus ASP_{fun} is functional because method update is realized on a copy of the active object: there are no side-effects.

1) ς-calculus: Objects consist of a set of labeled methods $[l_i = \varsigma(y)b]^{i\in 1..n}$. Attributes are considered as methods not using the parameters. The calculus features method call $t.l(s)$ and method update $t.l := \varsigma(y)b$ on objects where ς is the binder for the method parameter y. Every method may also contain a "*this*" element representing the surrounding object. Note, that the "this" is usually [1] expressed as an additional parameter x to each method's ς scope but we use for this exposition literally *this* to facilitate the understanding. It is, however, important to bear in mind that formally *this* is a variable representing a copy of the current object and that this variable is scoped as a local variable for each object. The ς-calculus is Turing complete, e.g. it can simulate the λ-calculus. We illustrate the ς-calculus by our example below.

2) Syntax of ASP$_{fun}$: ASP$_{\text{fun}}$ is a minimal extension of the ς-calculus by one single additional primitive, the *Active*, for creating an activity. In the syntax (see Table I) we distinguish between underlined constructs representing the static syntax that may be used by a programmer, while futures and active object references are created at runtime. We use the naming

$s, t ::= \underline{x}$	variable
$\mid \underline{this}$	generic object reference
$\mid \underline{[l_j = \varsigma(y_j)t_j]^{j\in 1..n}}$	$(\forall j,\ this \neq y_j)$ object definition
$\mid \underline{s.l_i(t)}$	$(i \in 1..n)$ method call
$\mid \underline{s.l_i := \varsigma(y)t}$	$(i \in 1..n, this \neq y)$ update
$\mid \underline{Active(s)}$	Active object creation
$\mid \alpha$	active object reference
$\mid f_i$	future

TABLE I. ASP$_{\text{FUN}}$ SYNTAX

convention s, t for ς-terms, α, β for active objects, f_k, f_j for futures, Q_α, Q_β for request queues.

3) Futures: A *future* can intuitively be described as a promise for the result of a method call. The concept of futures has been introduced in Multilisp [16] and enables asynchronous processing of method calls in distributed applications: on calling a method a future is immediately returned to the caller enabling the continuation of the computation at the caller side. Only if the method call's value is needed, a so-called wait-by-necessity may occur. Futures identify the results of asynchronous method invocations to an activity. Technically, we can see a future as a pair consisting of a future *reference* and a future *value*. The future reference points to the future value which is the instance of a method call in the request queue of a remote activity. In the following, we will use future and future *reference* synonymously for simplicity. Futures can be transmitted between activities. Thus different activities can use the same future.

4) Configuration: A *configuration* is a set of activities

$$C ::= \alpha_i[(f_j \mapsto s_j)^{j\in I_i}\ ,\ t_i]^{i\in 1..p}$$

where $\{I_i\}$ are disjoint subsets of $\mathbb{N}$. The unordered list $(f_j \mapsto s_j)^{j\in I_i}$ represents the request queue, t_i the active object, and $\alpha_i \in \mathrm{dom}(C)$ the activity reference. A configuration represents the "state"of a distributed system by the current parallel activities. Computation is now the state change induced by the evaluation of method calls in the request queues of the activities. Since ASP$_{\text{fun}}$ is functional, the *local* active object does not change – it is immutable – but the configuration is changed *globally* by the stepwise computation of requests and the creation of new activities.

The constructor $Active(t)$ activates the object t by creating a new activity in which the object t becomes active object. Although the active object of an activity is immutable, an update operation on activities is provided. It performs an update on a freshly created copy of the active object placing it into a new activity with empty request queue; the invoking context receives the new activity reference in return. If we want to model operations that change active objects, we can do so using the update. Although the changes are not literally performed on the original objects, a state change can thus be implemented at the level of configurations (for examples see [18]). Efficiency is not the goal of ASP$_{\text{fun}}$ rather minimality of representation with respect to the main decisive language features of active objects while being fully formal.

5) Results, Programs and Initial Configuration: A term is a result, i.e., a totally evaluated term, if it is either an object (like in [1]) or an activity reference. We consider results as values.

In a usual programming language, a programmer does not write configurations but usual programs invoking some distribution or concurrency primitives (in ASP$_{\text{fun}}$ *Active* is the only such primitive). This is reflected by the ASP$_{\text{fun}}$ syntax given above. A "program" is a term s_0 given by this static syntax (it has no future or active object reference and no free variable). In order to be evaluated, this program must be placed in an initial configuration. The initial configuration has a single activity with a single request consisting of the user program:

$$\mathit{initConf}(s_0) = \alpha[f_0 \mapsto s_0, []]$$

Sets of data that can be used as *values* are indispensable if we want to reason about information flows. In ASP$_{\text{fun}}$, such values can be represented as results (see above) to any configuration either by explicit use of some corresponding object terms or by appropriate extension of the initial configuration that leads to the set-up of a data base of basic datatypes, like integers or strings.

6) Informal Semantics of ASP$_{fun}$: Syntactically, ASP$_{\text{fun}}$ merely extends the ς-calculus by a second parameter for methods (the first being *this*) and the *Active* primitive but the latter gives rise to a completely new semantic layer for the evaluation of distributed activities in a configuration.

Local semantics (the relation $\rightarrow_\varsigma$) and the *parallel* (configuration) semantics (the relation $\rightarrow_\parallel$) are given by the set of reduction rules informally described as follows (see Appendix C for the formal semantics).

- CALL, UPDATE, LOCAL: the local reduction relation $\rightarrow_\varsigma$ is based on the ς-calculus.
- ACTIVE: $Active(t)$ creates a new activity α, with t as its active object, global new name α, and initially no futures; in ASP$_{\text{fun}}$ notation this is $\alpha[\varnothing, t]$.
- REQUEST, SELF-REQUEST: a *method call* $\beta.l(t)$ creates a new future f_k for the method l of active object β placing the resulting future value onto β's request queue; the future f_k can be used to refer to the future value $\beta.l(t)$ at any time.
- REPLY: *returns result*, i.e., replaces future f_k by the referenced result term, i.e., the future value resulting from some $\beta.l(t)$.

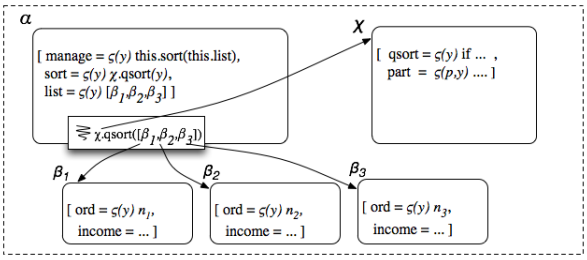


Fig. 3. Three active objects $\beta_1, \beta_2, \beta_3$ in controller α's list.

- UPDATE-AO: *active object update* creates a copy of the active object and updates the active object of the copy – the original remains the same (functional active objects are *immutable*).

C. Running Example: Private Sorting

As an example for a standard program consider the implementation of quicksort as an active object χ illustrated in Figure 3. The operations we use are :: for list cons, @ for list append, $\#$ for list length, hd for the list head, and a let construct (see [18] for details on their implementation).

$$
\begin{array}{l}
\chi\big[\varnothing, \\
\quad [\text{qsort} = \varsigma(y) \text{ if } y = [] \text{ then } [] \\
\qquad\qquad \text{else let } (a :: l) = y \\
\qquad\qquad\qquad (l_1, l_2) = \mathit{this}.\text{part } (a, l) \\
\qquad\qquad\qquad l_1' = \text{if } \#l_1 \leq 1 \text{ then } l_1 \text{ else } \mathit{this}.\text{qsort}(l_1) \\
\qquad\qquad\qquad l_2' = \text{if } \#l_2 \leq 1 \text{ then } l_2 \text{ else } \mathit{this}.\text{qsort}(l_2) \\
\qquad\qquad\quad \text{in } l_1' @ [a] @ l_2' \\
\qquad\qquad \text{end}, \\
\quad \text{part} = \varsigma(p, y) \text{ if } y = [] \text{ then } ([], []) \\
\qquad\qquad \text{else let } (a :: l) = y \\
\qquad\qquad\qquad (l_1, l_2) = \mathit{this}.\text{part } (p, l) \\
\qquad\qquad\quad \text{in if } p < a.\text{ord then } (l_1, a :: l_2) \text{ else } (a :: l_1, l_2) \\
\qquad\qquad \text{end} \\
\big]\big]
\end{array}
$$

The quick sort algorithm in χ is parametric over a method "ord", a numerical value, that is used in method "part". This method ord is assumed to be available uniformly in the target objects contained in the list that shall be sorted. We omit the parameter to calls of ord because it is unused, i.e., the empty object [].

The following controller object α holds a list of active objects (for example $[\beta_1, \beta_2, \beta_3]$ in Figure 3 but generally arbitrary thus represented as ... below). Controller α uses the quick sort algorithm provided by χ to sort this list on execution of the manage method.

$$
\begin{array}{l}
\alpha\big[\varnothing, [\text{manage} = \varsigma(y) \mathit{this}.\text{sort}(\mathit{this}.\text{list}), \\
\qquad\quad \text{sort} = \varsigma(y)\ \chi.\text{qsort}(y), \\
\qquad\quad \text{list} = \ldots]\big]
\end{array}
$$

The target objects contained in $\alpha's$ list (omitted) are active objects of the kind of β below. Here, the n in the body of method ord is an integer specific to β and the field income shall represent some private confidential data in β.

$$\beta\big[\varnothing, [\text{ord} = \varsigma(y)n, \text{income} = \ldots]\big]$$

If active objects of the kind of β represent principals in the system, it becomes clear what is the privacy challenge: the controller object α should be able to sort his list of β-principals without learning anything about their private data, here income.

III. Semi-Lattice Model for ASP_{FUN}

As a proof of concept, we show that the calculus of functional active objects ASP_{fun} gives rise to a fairly straightforward implementation of the security semi-lattice by mapping the concepts of the security model onto language concepts as follows.

- The global class ordering on sets of activity identities corresponds to the sets of activity references that are accessible from within an activity. We name this accessibility relation visibility (see Definition 3.1). It is a consequence of the structure of a configuration thereby at the discretion of the configuration programmer.
- The local classification of methods into public L and private H methods is specified as an additional security assignment mapping method names to $\{L, H\}$ at the discretion of the user.
- Based on these two devices for specifying and implementing a security policy with active objects we devise as a practical verification tool a security type system for ASP_{fun}. The types of this type system correspond quite closely to the security classes of the semi-lattice defined in Section II-A: object types are pairs of security assignment maps and global levels.

A. Assigning Security Classes to Active Objects

Visibility: We define visibility as the "distributed part" of the accessibility within a configuration. It derives from the activity references and thus represents the global security specification as programmed into a configuration.

Definition 3.1 (Visibility): Let C be a configuration with a security specification *sec* partitioning the methods of each of C's active objects locally into H and L methods. Then, the relation $\leq_{VI}$ is inductively defined on activity references by the following two cases.

$$
\begin{array}{rcl}
\left(\begin{array}{c} \beta[Q_\beta, [l_i = \varsigma(y)t_i]^{i\in 1..n}] \in C \\ \mathit{sec}(l_i) = L \wedge t_i = E[\alpha] \end{array}\right) & \Rightarrow & \alpha \leq_{VI} \beta \\
\left(\begin{array}{c} \beta[Q_\beta, [l_i = \varsigma(y)t_i]^{i\in 1..n}] \in C \\ \mathit{sec}(l_i) = L \wedge t_i = E[\gamma] \wedge \alpha \leq_{VI} \gamma \end{array}\right) & \Rightarrow & \alpha \leq_{VI} \beta
\end{array}
$$

We use the vertical notation $\binom{\phi}{\xi}$ to abbreviate $\phi \wedge \xi$; for context variable E see Appendix C. We then define the relation called *visibility* $\sqsubseteq_C^{sec}$ as the *reflexive transitive closure* over $\leq_{VI}$ for any C, *sec*. □

We denote the *visibility range* using Definition 3.1 as $VI_{sec}(\alpha, C) \equiv \{\beta \in \text{dom}(C) \mid \beta \sqsubseteq_C^{sec} \alpha\}$. The visibility relation extends naturally to a relation $\sqsubseteq$ on global levels: every activity $\alpha \in C$ may be assigned the global level corresponding to the union of all its visible activities $VI_{sec}(\alpha, C)$. This relation is a subrelation of the subset relation on the powerset of activity identities introduced before and thus also a partial order. We use it as the semantics of the subtype relation in Section IV.

Assigning Security Classes to Example: To illustrate how activities are labeled in the semi-lattice model using visibility, consider the running example above where we assume the list in controller α to contain various active object references

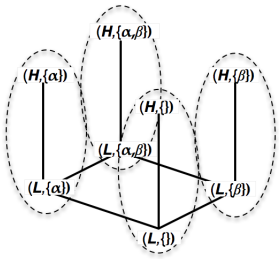


Fig. 4. Tentatively drawing in object classes as confinement zones.

$[\beta_0, \ldots, \beta_n]$. We assign to each activity the global class containing its own identity and those of all its visible activities. For our example, the *global* class of controller α would be the following.

$$\delta_\alpha = \{\alpha\} \cup \delta_\chi \cup \bigcup_{i=0..n} \delta_{\beta_i}$$

The global classes δ_{β_i} of the β_i objects and δ_χ in turn contain all their visible objects' classes. Thus, the global classes are ordered $\delta_{\beta_i} \subseteq \delta_\alpha$ for all i and $\delta_\chi \subseteq \delta_\alpha$. The security classification of methods assigns pairs of global classes and local levels to method names, for example, $\mathrm{ord}_{\beta_i} \mapsto (L, \delta_{\beta_i})$ and $\mathrm{income}_{\beta_i} \mapsto (H, \delta_{\beta_i})$.

Practical Classification of Objects: The pairs (S, δ) in the partial order *CL* (see Section II-A3) are the security classes for methods of active objects. The semi-lattice is actually defined as a partial order on object methods rather than objects. To classify objects we consider only the global part of the classification, i.e., the second δ component because all methods of an active object have this δ in common. Intuitively, this factorization corresponds to drawing objects as borders into the semi-lattice structure (see Figure 4). These borders represent the confinement zone of an active object.

Formally, we consider an object class to be the factorization $([l_i \mapsto S_i], \delta)$: a pair of a security assignment to $\{L, H\}$ for each method l_i of an object and the object's global class δ common for all parts. An activity contains one active object but may contain various passive objects. The security assignment of an active object must be defined for all contained objects (see rule SECASS SUBSUMPTION in Section IV).

B. Secure Down Calls

In a distributed system with a nontrivial security classification of communicating objects, secure down calls are not possible because they would violate the security policy of "no-down-flows" of information. In general, a method call represents an information flow to the remote object in the form of the request itself and the parameters passed; its response flows information back in the form of a reply. Therefore, secure method communication is trivially restricted to objects of one class – otherwise one direction would contradict the policy "no-down-flows". This catch-22 situation can be overcome if we exclude side-effects: the requests do not leave traces in the remote object. In $\mathrm{ASP}_{\mathrm{fun}}$ this is given implicitly by the semantics because requests created by method calls in the remote object are not accessible by the remote object itself. However, the reply may flow information up. Thus, information does flow back, i.e. up.

As an overall result of the properties presented in this paper we can infer that secure down calls are possible. The reasoning is as follows. We assume as given a configuration together with a security specification *sec* partitioning a portion of the methods into public (L) and private (H). If this configuration can be type checked according to our type system, it is secure, i.e., we know it has confinement and is noninterfering as we are going to see in Section V. Therefore, futures can be securely used in higher security classes, i.e., method results may flow up but, since no implicit flows exist, information is not leaked in the process.

Side-effect freedom does permit to securely call down because the call leaves no visible trace. But does this not also exclude any mutual information exchange on the same level? It might seem so, but fortunately, if we have two activities that are in the same class, methods calls between them are possible permitting bidirectional information flow. As an example, an implementation of the Needham-Schroeder public key protocol is given next.

Needham Schroeder Public Key Protocol (NSPK) in ASP_{fun}

This example illustrates that inside one security class mutual information exchange is possible between different activities. The easiest way to illustrate this is to use a protocol. We use the corrected short form of the Needham Schroeder Public Key Protocol (NSPK) originally published by [28]. The originally published protocol missed out the B inside the encrypted message to A in step two thereby giving rise to the well-known attack of [22].

The protocol is usually written as follows using public keys K_A, K_B known globally and their secret counterparts K_A^{-1}, K_B^{-1} establishing nonces N_A, N_B in the process of authentication.

$$\begin{array}{lcl} A \to B & : & \{N_A, A\}_{K_B} \\ B \to A & : & \{N_A, N_B, B\}_{K_A} \\ A \to B & : & \{N_B\}_{K_B} \end{array}$$

In $\mathrm{ASP}_{\mathrm{fun}}$, the protocol is implemented as a set of methods between two activities A and B. We omit details about decoding and keys because it is clear that they can be implemented and we want to highlight the communication process.

$A = [\varnothing,$
$[\mathrm{own}_{id} = \ldots$
$\mathrm{B}_{id} = \ldots$
$\mathrm{step}_1 = \varsigma(y)$
  let $N_A = \mathrm{new\_nonce}$
    $\mathit{reply} = B.\mathrm{step}_2(\{N_A, \mathit{this}.\mathrm{own}_{id}\}_{K_B})$
    $(N'_A, N'_B, B') = K_A^{-1}(\mathit{reply})$
  in if $B' = \mathit{this}.B_{id} \wedge N'_A = N_A$
    then $(\mathit{this}.\mathrm{knows} := N_B).\mathrm{NA} := N_A$
    else $\mathit{this}.\mathrm{knows} := \mathrm{error}$,
$\mathrm{step}_3 = \varsigma(y)$
  let $(N'_A, N_B, B'_{id}) = \{y\}_{K_A^{-1}}$
    if $N'_A = \mathit{this}.\mathrm{NA} \wedge B'_{id} = \mathit{this}.B_{id}$
    then $\{N_B\}_{K_B}$
    else $\mathit{this}.\mathrm{knows} := \mathrm{error}$,
$\mathrm{knows} = \ldots,$
$\mathrm{NA} = \ldots]]$

The protocol can be executed by invoking method $A.\mathrm{step}_1$

which in turn invokes the step $B.\text{step}_2$ and $A.\text{step}_3$.

$B = [\varnothing,$
$[\text{own}_{id} = \ldots$
$\text{A}_{id} = \ldots$
$\text{step}_2 = \varsigma(y)$
let $N_B = \text{new\_nonce}$
$(N'_A, A'_{id}) = \{y\}_{K_B^{-1}}$
in if *this*.$A_{id} = A'_{id}$
then let *reply* $= A.\text{step}_3(\{N'_A, N_B, \textit{this}.\text{own}_{id}\}_{K_A})$
$N'_B = \{\textit{reply}\}_{K_B^{-1}}$
in if $N'_B = N_B$
then (this.knows $:= N_A$).NB $:= N_B$
else *this*.knows := error,
else *this*.knows := error,
$\text{knows} = \ldots,$
$\text{NB} = \ldots]]$

In each of the steps the nonces are created, encrypted and tested between the method calls. If the communicated messages adhere to the protocol, i.e., the nonces and ids correspond to what has been sent in earlier steps, the own nonces are updated into the methods A.NA and B.NB and the other's nonces in the respective method "knows". Otherwise, the protocol failure is recorded as "error" in method knows. This protocol implementation illustrates that mutual information flows are possible locally within one security class. The type system that we present in the following Section IV accepts this configuration since the calls are of the same global level δ.

C. Security Analysis

In language based security, we may use the means provided by a language to enforce security. That is, we make use of certain security guarantees that correspond to implicit assumptions concerning the execution of programs. The language introduces a security perimeter because we assume that the language compilation and run-time system are respected (below the perimeter) while the language is responsible for the security above the perimeter by virtue of its semantics and other language tools, e.g. static analysis by type checking. We now describe the security goal of confidentiality addressed in this paper and elaborate on the attacker model for active objects.

Security Goal Confidentiality: A computation of active objects is an evaluation of a distributed set of mutually referencing activities. Principals that use the system can observe the system only by using the system's devices. We make the simplifying assumption that principals can be identified as activities. Principals, objects, programs and values are thus all contained in this configuration. There are no external inputs to this system – it is a closed system of communicating actors. We concentrate in this paper on confidentiality, i.e. activities should not learn anything about private parts of other activities neither directly nor indirectly. Integrity is the dual to this notion and we believe that it can simply be derived from our present work by inverting the order relation.

Attacker Model: As a further consequence to the language based approach to security, we restrict the attacker to only have the means of the language to make his observations. Consequently, we also consider the attacker – as any other principal – as being represented by an activity. The attacker's knowledge is determined by all active objects he sees, more precisely their public parts. If any of the internal computations in inaccessible parts of other objects leak information, the attacker can learn about them by noticing differences in different runs of the same configuration. Inaccessible parts of other objects must be their private methods or other objects that are referenced in these private parts. The language semantics and the additional static analysis must guarantee that under the assumption of the security perimeter an attacker cannot learn anything about private parts.

D. Information Flow Control

Information flow control [11] technically uses an *information flow policy* which is given by the specification of a set of *security classes* to classify information and a *flow relation* on these classes that defines allowed information flows. System entities that contain information, for example variables x, y, are bound to security classes. Any operation that uses the value of x to calculate that of y, creates a flow of information from x to y. This operation is only admissible if the class of y dominates the class of x in the flow relation, formally written $\delta_x \sqsubseteq \delta_y$ where δ_e denotes the class of entity e. The concept of information flow classically stipulates that the security classes together with the flow relation as an order relation on the classes are a lattice [10], [8]. We differ here since we only require a semi-lattice.

Information Flow Control for Active Objects: Information is contained in data values which are here either objects or activity references (see Section II-B). To apply the concept of information flow control to configurations of active objects, we need to interpret the above notions of security classes, their flow relation, and the entities that are assigned to the security classes: we identify the classes of our security model as the security classes of methods and the flow relation as the semi-lattice ordering on these classes (see Section II-A). Flows of information local to objects are generated by local method calls between neighboring methods of the same object. These are regulated by the local L/H-classification of an object's methods (H may call L *and* H – but L only L). Global flows result from remote method calls between objects' methods. The combined admissible flows have to be in accordance with a concrete configuration and its L/H specification.

E. Enforcing Legal Information Flows

To illustrate the task of controlling information flows, we first extend the intuition about information flow to configurations of active objects. An active object sees only other active objects that are directly referenced in its methods or those active objects that are indirectly visible via public methods of visible objects. From the viewpoint of one active object, information may flow into the object and out of the object. For each direction, there are two ways how information may flow: implicit or explicit (direct) flows. Information flows *explicitly into* an object by parameters passed to remote calls directed to the object's methods; it may also flow *implicitly into* the object simply if the choice of which method is called depends on the control flow of the calling object. Similarly, information flows *explicitly out* of an active object by parameters passed to remote method calls and *implicitly out* of it, if the choice depends on the object's own control flow. Some of these flows are illustrated on our running example next.

Running Example: Implicit Information Flow: We will now finally illustrate the security model on the running example showing implicit information flows of active objects introduced above in Section II-C. Let us assume that the implementation of the β-objects featuring in the controller's list had the following implementation.

$$\beta\big[\varnothing,\quad \begin{array}{l}[\text{ord} = \varsigma(y) \text{ if } \textit{this}.\text{income}/10^3 \geq 1 \text{ then } 1 \text{ else } 0],\\ \text{income} = \ldots \quad]\end{array}$$

Let us further assume that ord is again a public method and income again the private field of β. We have here a case of an implicit information flow. Since the guard of the if-command in ord depends on the private field income, effectively the order number of a β-object is 1 if the income of β is more than 1000 else 0. In our security model this control flow represents an illicit flow of information from a high level value in β to its public parts and is thus visible to the remote controller. This should not be the case since $H \sqsupset \text{L}$. It should thus be detectable by an information flow control analysis. We will show next how to detect it statically by a security type system.

IV. Security Type System

Before formalizing security of active objects and defining a type system that implements rules for a static analysis, we summarize the security considerations so far and motivate the upcoming type system and proofs.

A. Intermediate Summary, Motivation, and Outlook

In a configuration of active objects we may have direct (explicit) and implicit information flows through method calls which are controlled differently.

- To guarantee only legal information flows on direct calls we rely on the labeling of methods by L and H and on the global hierarchy. This corresponds to the simple security property of *confinement*: remote method calls can refer only to low methods of visible objects. Confinement can be locally checked. It is decidable since it corresponds to merely looking up method labels in a security assignment.
- We will use a program counter PC that records the current security level of a method evaluation. Locally, within the confinement zone of an activity, accessing H-methods in L-methods may create implicit flows – as seen in the example. To detect such flows and protect the confidential information from flowing out of the confinement zone of the activity, the program counter records these dependencies by increasing to H. In combination with the method labels, the PC thereby allows associating the calling context with the called method. Implemented into type rules, this enables static checking and thus controlling of information flows in evaluations of configurations.

As a security enforcement mechanism of our multilateral security model for active objects, we propose a security type system, i.e., a rule set for static analysis of a configuration with respect to its methods' security assignment. The idea of a security type system is as follows. Not all possible programs in ASP_{fun} are secure. In general, for example, any method can be accessed in an active object. The purpose of a type system for security is to supply a set of simple rules defining types of configurations enabling a static check (before runtime) whether those contain only allowed information flows.

The above described cases of information flows need to be implemented in the type rules such that the rules allow to infer a type just for secure configurations and otherwise reject them. The first direct case of information flow is intuitively simple, as it boils down to locally looking-up the security level of a method before deciding whether a remote call from up in the hierarchy can be granted. The "up in the hierarchy" is encoded in a subtype relation $\sqsubseteq$ encoding the global hierarchy described by the visibility relation. After the presentation of the type system in this section, we prove in the following Section V that confinement is a security property implied by it.

How to avoid and detect implicit flows, is more subtle: the combination of a program counter PC with the called method's security label grants us to combine the provenance of one call with the security level of the call context. However, this combination needs to adhere to the security specification for all runs of a program and thus all possible calls in a context. The appropriate notion of security for this is noninterference: in all runs the observable (low) parts of configurations need to look "the same". Therefore, we first introduce a notion of noninterference for active objects based on which we will then be able to express the absence of implicit flows and prove multilateral security. The definition of noninterference and proofs of properties are contained in Section V. We first introduce the type system.

B. Type System

Type Formation: We need to provide types for objects and for configurations of active objects; the latter by mapping names of futures and activities to object types. The two-dimensional classification of local and global security described above translates directly into the object types of the security type system. A type is a pair $([l_i \mapsto S_i]^{i=1..n}, \delta_\alpha)$ where $S_i \in \{L, H\}$. The first part $[l_i \mapsto S_i]^{i=1..n}$ provides the partition of methods into public (L) and private (H) methods for the object. The other element δ_α of an object type represents the global classification of an object. This global level corresponds to the classification of the object's surrounding activity α derived from its visibility. We adopt the following naming conventions for variables. δ stands for the global part of a type. We use A to denote security assignments, e.g. $A = [l_i \mapsto S_i]^{i=1..n}$. S_i, or simply S, stands for levels L or H. In general, we use indexed variables to designate result values of a function, e.g., S_i for the level value of method l_i – also expressed as $A(l_i)$. We use Σ for object types $\Sigma = ([l_i \mapsto S_i]^{i=1..n}, \delta)$. To map an object type Σ to its security assignment or its global part, respectively, we use the projections $ass(\Sigma)$ and $glob(\Sigma)$. We formally use a parameter *sec* as the parameter for the overall methods' security assignment of an entire configuration C. A triplet of maps is a configuration type $\langle \Gamma_{act}, \Gamma_{fut}, sec\rangle$ assigning types to all activities and futures of a configuration in addition containing the security assignment *sec*.

Typing Relations: A typing judgement $T; S \vdash x : ([l_i \mapsto S_i]^{i=1..n}, \delta)$ reads: given type assumptions in T, term x has type $([l_i \mapsto S_i]^{i=1..n}, \delta)$ in the context of a program counter $S \in \{L, H\}$. A program counter (PC) is a common technique in information flow control originating in Fenton's Data Mark Machine [12]. The PC encodes the highest security level that has been reached in all possible control flows leading to the current control state. In a functional language, like ASP_{fun}, this highest security level of all execution paths simply is the level of the evaluation context for the term x. Thus, the PC is some $S \in \{L, H\}$ denoting the security label of the local context. The type environment T contains types Σ for the parameter variables y and types for the parameter *this* both paired with the local security level S representing their local PC.

Subsumption Rules: Subsumption means that an element of a type also has the type of its super-type. It is responsible for making the partial order relation on global levels a subtype relation. Intuitively, GLOB SUBSUMPTION says that if a term can be typed in a low context it may as well be "lifted", i.e., considered as of higher global level thereby enforcing (together with TYPE CALL below) that only L-methods can be accessed remotely. This corresponds to the confinement property as formally shown in Section V-B. The local security class ordering is $L \leq H$ and features implicitly in the type system in the form of a second – the local – subsumption rule. Finally, the rule SECASS SUBSUMPTION allows the security assignment type of an object to be extended. This rule is necessary to consider an object also as a local object inside another (active) object adopting its security assignment.

$$\text{LOC SUBSUMPTION}\quad \frac{T; L \vdash x : (A, \delta)}{T; H \vdash x : (A, \delta)} \qquad \text{SECASS SUBSUMPTION}\quad \frac{T; S \vdash x : (A, \delta) \qquad A \subseteq A'}{T; S \vdash x : (A', \delta)}$$

$$\text{GLOB SUBSUMPTION}\quad \frac{T; L \vdash x : (A, \delta) \qquad \delta \sqsubseteq \delta'}{T; L \vdash x : (A, \delta')}$$

TABLE II. SUBSUMPTION RULES, $A = [l_i \mapsto S_i]^{i \in 1..n}$, $S \in \{L, H\}$

$$\text{VAL SELF}\quad \textit{this} : \Sigma :: T; \sqcup_{i \in 1..n} S_i \vdash \textit{this} : \Sigma \qquad \text{VAL LOCAL}\quad x : \Sigma :: T; S \vdash x : \Sigma$$

$$\text{TYPE OBJECT}\quad \frac{\forall i \in 1..n.\ \ \textit{this} : \Sigma :: y : \Sigma :: T; S_i \vdash t_i : \Sigma}{T; \sqcup_{i \in 1..n} S_i \vdash [l_i = \varsigma(y) t_i]^{i \in 1..n} : \Sigma} \qquad \text{TYPE CALL}\quad \frac{T; S \vdash o : \Sigma \qquad j \in 1..n \qquad T; S_j \vdash t : \Sigma}{T; S_j \vdash o.l_j(t) : \Sigma}$$

$$\text{TYPE UPDATE}\quad \frac{T; S \vdash o : \Sigma \qquad j \in 1..n \qquad \textit{this} : \Sigma :: y : \Sigma :: T; S_j \vdash t : \Sigma}{T; S \vdash o.l_j := \varsigma(y) t : \Sigma}$$

TABLE III. TYPE RULES FOR OBJECTS; $\Sigma = ([l_i \mapsto S_i]^{i \in 1..n}, \delta)$

Object Typing: The object typing rules in Table III describe how object types are derived for all possible terms of ASP_{fun}. The VAL-rules state that type assumptions stacked on the type environment T left of the turnstile $\vdash$ can be used in type judgments. These rules apply to the two kinds of environment entries for *this* and for the y-parameter. Since the *this* represents the entire object value itself, its PC is derived as the supremum of all security levels assigned to methods in it. We express this supremum as the join over all levels $\sqcup_{i \in 1..n} S_i$. The other rules are explained as follows. TYPE OBJECT: if every method l_i of an object is typeable with some local type $S_i \in \{L, H\}$ assigned to it by the assignment component A of Σ, then the object comprising these methods is typeable with their maximal local type. Thus, objects that contain H methods cannot themselves be contained in other L-methods. Otherwise, local objects containing confidential parts could be typed with GLOB SUBSUMPTION at higher levels (see the Appendix for a "borderline example" illustrating this point). Only objects that are purely made from L-methods can be accessed remotely in their entirety. Albeit this strong restriction, the CALL rule permits selectively accessing L methods of such objects (see below). The *PC* guarantees that all method bodies t_i are typeable on their given privacy level S_i. The rule TYPE CALL is the central rule enforcing that only L methods can be called in any object – locally or remotely. Initially, a call $o.l_j(t)$ can only be typed as $\Sigma = ([l_i \mapsto S_i]^{1=1..n}, \delta)$ for the δ of the surrounding object o. Although the PC in the typing of o is (by TYPE OBJECT) the maximal level of all methods, we may still call L-methods on objects that are typed with PC as H. The PC in the typing of the resulting call $o.l_j(t)$ is coerced to S_j, i.e., the security level assigned to the called method. This prevents H methods from being callable remotely while admitting to call methods on objects that are themselves typed in a *H-PC*. Because of the rule GLOB SUBSUMPTION any method call $o.l_j(t) : (A, \delta)$ can also be interpreted as $o.l_j(t) : (A, \delta')$ for $\delta \sqsubseteq \delta'$ but this is restricted to L contexts: a method call typeable in an H context cannot be "lifted", i.e., it cannot be interpreted as well-typed with δ'; to prevent this, the *PC* in GLOB SUBSUMPTION is L thus excluding CALL instantiations for methods l_j with $S_j = H$. UPDATE: an update of an object's method is possible but conservatively, i.e., the types must remain the same.

Configuration Typing: The rules for configurations (see Table IV) use the union of all futures of a configuration.

Definition 4.1 (Future Domain): Let C be a configuration. We define the domain of all futures of C.

$$\mathit{dom}_{\mathit{fut}}(C) \equiv \bigcup \{\text{dom}(Q) \mid \exists\ \alpha, a.\ \alpha[Q, a] \in C\} \qquad \square$$

The rules for configurations anticipate two semantic properties of futures in well formed ASP_{fun} configurations. We use well-formedness of ASP_{fun} configurations as defined in [18]; in brief: there are no dangling references.

Property 4.2 (Unique Future Home Activity): Every future is defined in the request queue of one unique activity.

$$\forall f_k \in \text{dom}_{\mathit{fut}}(C).\ \exists! \alpha[Q, a] \in C.\ f_k \in \text{dom}(Q)$$

We denote this unique activity α as $\mathit{futact}_C(f_k)$. $\square$

Next, every future f_k in a well formed configuration C is created by a call to a unique label in its home activity.

Property 4.3 (Unique Future Label): Let $\alpha[Q, a] \in C$ be the unique $\mathit{futlab}_C(f_k)$. Then,

$$\forall f_k \in \text{dom}_{\mathit{fut}}(C).\ \exists! l \in \text{dom}(a).\ \exists t.\ a.l(t) \rightarrow_{\|}^{*} Q(f_k).$$

We denote this unique method label as $\mathit{futlab}_C(f_k)$. $\square$

We omit the configuration C for the previous two operators if it is clear from context.

The configuration type rules link up types for activities and futures with the local types of terms in active objects and request lists (see Table IV).

$$\textsc{Type Active}\ \frac{\langle \Gamma_{act}, \Gamma_{fut}, sec\rangle, T; S \vdash a : \Sigma}{\langle \Gamma_{act}, \Gamma_{fut}, sec\rangle, T; S \vdash \mathit{Active}(a) : \Sigma} \qquad \textsc{Type Active Object Reference}\ \frac{\beta \in \mathrm{dom}(\Gamma_{act})}{\langle \Gamma_{act}, \Gamma_{fut}, sec\rangle, T, M_\beta \vdash \beta : \Gamma_{act}(\beta)}$$

$$\textsc{Type Future Reference}\ \frac{f_k \in \mathrm{dom}(\Gamma_{fut})}{\langle \Gamma_{act}, \Gamma_{fut}, sec\rangle, T; ass(\Gamma_{fut}(f_k))(futlab(f_k)) \vdash f_k : \Gamma_{fut}(f_k)}$$

$$\textsc{Type Configuration}\ \frac{\mathrm{dom}(\Gamma_{act}) = \mathrm{dom}(C) \qquad \mathrm{dom}(\Gamma_{fut}) = \mathrm{dom}_{fut}(C) \qquad \bigcup_{\alpha\in\mathrm{dom}(C)} A_\alpha \subseteq sec}{\forall\, \alpha[Q,a] \in C.\ \begin{cases} \langle \Gamma_{act}, \Gamma_{fut}, sec\rangle, \varnothing; M_\alpha \vdash a : \Gamma_{act}(\alpha) \wedge \\ \forall\, f_k \in \mathrm{dom}(Q).\ \begin{cases} \Gamma_{act}(\alpha) = \Gamma_{fut}(f_k) \wedge \\ \langle \Gamma_{act}, \Gamma_{fut}, sec\rangle, \varnothing; A_\alpha(futlab(f_k)) \vdash Q(f_k) : \Gamma_{fut}(f_k) \end{cases} \end{cases}}{\vdash C : \langle \Gamma_{act}, \Gamma_{fut}, sec\rangle}$$

TABLE IV. TYPING CONFIGURATIONS; $M_\alpha = \bigsqcup_{i\in 1..n_\alpha} A_\alpha(i)$; $A_\alpha = ass(\Gamma_{act}(\alpha))$, WHERE $ass([l_i \mapsto S_i]^{i=1..n}, delta) = [l_i \mapsto S_i]^{i=1..n}$.

TYPE ACTIVE allows to transfer the type of an object term to its activation which coerces the types of activities and activity references to coincide with the types of their defining objects. This is achieved together with TYPE ACTIVE OBJECT REFERENCE and the clause $\langle \Gamma_{act}, \Gamma_{fut}, sec\rangle, \varnothing \vdash a : \Gamma_{act}(\alpha)$ of TYPE CONFIGURATION.

TYPE FUTURE REFERENCE similarly assigns the types for the future references in Γ_{fut}. For a given activity α, this rule further coerces the PC for the typing of f_k to coincide with $A_\alpha(futlab(f_k))$, i.e., α's security assignment applied to the label that leads to the instance of f_k.

The rule TYPE CONFIGURATION ensures consistency between the type maps Γ_{fut}, Γ_{act}, and the overall security assignment *sec*. It looks rather complex but it essentially only scoops up what has been prepared by the other rules. The first two clauses ensure that the domains of activities coincide with the configuration domain and similarly for futures that the future type map Γ_{fut} is defined over precisely all futures in all activities. The third clause integrates the security specification *sec* to be respected by the individual security assignments of activities. The last large clause of TYPE CONFIGURATION specifies first that the activity types assigned to activity references by Γ_{act} coincide with their active object types. The second part of that clause addresses the future types in Γ_{fut}. Note, that in the context of this clause we may assume $\alpha = futact(f_k)$ by Property 4.2. The clause ensures that the types assigned by Γ_{fut} coincide with the ones assigned by Γ_{act} to their home activity. Additionally, this final clause ensures that the request $Q(f_k)$ must have the type assigned by the future map Γ_{fut} for this future f_k with the PC that corresponds to the PC assigned by the security assignment in the home activity.

C. Running Example: Type System Checks Example

For the sake of argument, we illustrate the application of the type system with an inconsistent constraint on the assignment *sec* for the example in Section II-C. The extended implementation as discussed in Section III-E contains the following changed ord function (we repeat the code here for convenience).

$$\beta[\varnothing,\quad [\mathrm{ord} = \varsigma(y)\ \text{if } \mathit{this}.\mathrm{income}/10^3 \geq 1 \text{ then } 1 \text{ else } 0],\ \mathrm{income} = \ldots\quad]$$

If we specify income as private, this extended version of the running example may contain an implicit illegal information flow. Any security assignment *sec* that fulfills the constraint must be fallacious since the call β_i.ord in the manager object α reveals information about the confidential (H) value of income. The type system rejects any such *sec* since no consistent type can be inferred for the configuration in this case as we illustrate next. The failed type checking thus proves that for the extended configuration all specifications would have to specify income $\mapsto L$ because the assumption income $\mapsto H$ was inconsistent.

The global classification is derived according to the visibility relation (Definition 3.1) from the example's configuration as $\delta_{\beta_i} \sqsubseteq \delta_\alpha$ for all i. To be able to type the call to β_i.ord in manager object α this method must be an L-method according to TYPE CALL. Hence, we need to have the following extended constraint on *sec*.

$$sec \supseteq \quad \{\mathrm{ord} \mapsto L, \mathrm{income} \mapsto H\}$$

The third clause of TYPE CONFIGURATION, i.e., $\bigcup_{\alpha\in\mathrm{dom}(C)} ass(\Gamma_{act}(\alpha)) \subseteq sec$, gives us the constraint $ass(\Gamma_{act}(\beta_i)) \supseteq \{\mathrm{ord} \mapsto L, \mathrm{income} \mapsto H\}$ since *sec* and $ass(\Gamma_{act}(\beta_i))$ are both functions.

We show now that β_i (for an arbitrary i in the configuration) cannot be typed with this type constraint. The final step in a type inference to arrive at a type $\Gamma_{act}(\beta_i)$ for β_i can only be an instance of TYPE OBJECT which looks as follows.

$$\textsc{Instance Type Object}\ \frac{\mathit{this} : \Sigma_{\beta_i} :: [] : \Sigma_{\beta_i} :: \varnothing; A(\mathrm{ord}) \vdash t_{\mathrm{ord}} : \Sigma_{\beta_i} \qquad \mathit{this} : \Sigma_{\beta_i} :: [] : \Sigma_{\beta_i} :: \varnothing; A(\mathrm{income}) \vdash t_{\mathrm{income}} : \Sigma_{\beta_i}}{\varnothing; \sqcup\{A(\mathrm{ord}), A(\mathrm{income})\} \vdash [\mathrm{ord} = \varsigma(y)\ t_{\mathrm{ord}}, \mathrm{income} = \varsigma(y)\ t_{\mathrm{income}}] : \Sigma_{\beta_i}}$$

We write Σ_{β_i} for $\Gamma_{act}(\beta_i)$, $A = sec(\Sigma_{\beta_i})$, and $t_{\mathrm{ord}} =$ if $\mathit{this}.\mathrm{income}/10^3 \geq 1$ then 1 else 0. In fact, a more technical definition of t_{ord} is

$$t_{\mathrm{ord}} = \quad (((\mathit{true}.\mathrm{if} := (\mathit{this}.> 0(\mathit{this}.\mathrm{div}_{10^3}(\mathit{this}.\mathrm{income})))).\mathrm{then} := 1).\mathrm{else}{:=}\ 0).\mathrm{if}([]).$$

where *true* is a boolean object containing methods if, then, and else. The details of this boolean object and its typing as well as the details of the following abridged reasoning are contained in Appendices A and B. The main point that we can see from this implementation is that the type $A(\mathrm{ord})$ is coerced by the type $A(\mathrm{if})$, i.e., it must hold that $A(\mathrm{ord}) = A(\mathrm{if})$ in Σ_{β_i}. This is the case, because t_{ord} is a call to the method if. According to the rule TYPE CALL, the PC must thus be S_{if} which corresponds here to $A(\mathrm{if})$ and coincides with the PC $A(\mathrm{ord})$ in the above instance of TYPE OBJECT, i.e., $A(\mathrm{ord}) = A(\mathrm{if})$. Now, the remaining argument just shows that $A(\mathrm{if})$ must be H. In short form, the reasoning for the latter goes as follows. By assumption, $A(\mathrm{income})$ must be H. Thus according to TYPE CALL and VAL SELF, *this*.income is typeable only with PC as H. We must apply TYPE CALL twice, to type $\mathit{this}.> 0(\mathit{this}.\mathrm{div}_{10^3}(\mathit{this}.\mathrm{income}))$. The PC for typing this is H each time because it must be the same as

the PC (named S_j) in typing the parameter (named t in the rule TYPE CALL) and the previous typing of the parameter *this*.income has a H-PC. The PC in the application of the rule TYPE UPDATE is then also coerced to H in the typing of the newly inserted body method l_j, here "if". Hence, this update coerces the PC S_{if} to be H, i.e., $A(\text{if})$ to be H. The two following updates do not change the security type of the method if. We are finished since as we have seen above $A(\text{if}) = A(\text{ord})$. Thus, $A(\text{ord})$ must be H and cannot be typed L as would be necessary to call this method remotely in α. The typing fails. We have a contradiction to the initially required specification that income be private. Since this was the only assumption, if follows by contraposition that income must be L to make the configuration typeable.

This illustrates the correctness of the type system by example: the configuration C of our running example cannot be typed with the constraint income $\mapsto H$ since any attempt to infer a type $\langle \Gamma_{act}, \Gamma_{fut}, sec \rangle$ for it fails. The type inference reveals the dependency between ord and income: a security leak because it would enable implicit information flows from β_i's private part to χ.

In the following, we provide general proofs showing that the type system is sound, i.e., it generally implies security not just for the example.

V. PROPERTIES

A. Preservation

Type safety includes always a preservation theorem: if a program can be typed, the type has to be preserved by the evaluation of the program – otherwise the guarantees encoded in the types would be lost. In our case, since configurations dynamically change during the evaluation with the reduction relation $\rightarrow_{\|}$, the preservation has a slightly unusual form as the configuration type actually changes. But this change is conservative, i.e., dynamically created new elements are assigned new types but old types persist, as represented below by $\subseteq$. Alongside the configuration types, also the security class lattice is extended likewise in a conservative way by extension of the visibility relation.

Theorem 1 (Preservation):

$$\left(\begin{array}{l} \vdash C : \langle \Gamma_{act}, \Gamma_{fut}, sec \rangle \\ C \rightarrow_{\|} C' \end{array} \right) \Rightarrow \exists\, \Gamma'_{act}, \Gamma'_{fut}.\ \vdash C' : \langle \Gamma'_{act}, \Gamma'_{fut}, sec \rangle$$

where $\Gamma_{act} \subseteq \Gamma'_{act}$ and $\Gamma_{fut} \subseteq \Gamma'_{fut}$.

The proof of this theorem has two parts. The first part shows a local preservation property for the part of the type system that describes secure method calls at the level of objects, i.e., the rules depicted in Tables II and III. The second part of the proof addresses the typing rules at the global level, i.e., the configuration typing rules depicted in Table IV. Both proofs are straightforward using the induction schemes corresponding to the inductive rule definitions of the type rule definitions. Albeit the relatively small size of the computation model ASP_{fun}, these rules are fairly complex. Hence to avoid mistakes in these proofs we have formalized them in Isabelle/HOL. The Isabelle/HOL sources can be found at https://sites.google.com/site/floriankammueller/home/resources.

B. Confinement

Confinement is the property of our type system encoding the principal idea of the security model: if a method of an object can be called remotely, it must be a public L method. As a preparation to proving confinement, we present next a chain of lemmas that lead up to it. Let o be an object and T be an arbitrary type environment throughout the following formal statements.

The type rules for subsumption allow that types of objects can be "lifted", i.e., objects can have more than one type. We lose uniqueness of type judgments. To overcome this, we use a well-known trick (already been used in the Hindley-Milner type system for ML to accommodate polymorphic types) to regain some kind of uniqueness: minimal types.

Definition 5.1 (Minimal Type): Define the minimal type in the PC context of $S \in \{L, H\}$ as follows.

$$T; S \vdash_{\text{ML}} o : (A, \delta) \equiv \left\{ \begin{array}{l} T; S \vdash o : (A, \delta) \wedge \\ \forall S', A', \delta'. \\ \\ T; S' \vdash o : (A', \delta') \Rightarrow \left(\begin{array}{c} S \leq S' \\ \delta \sqsubseteq \delta' \\ A \subseteq A' \end{array} \right) \end{array} \right.$$

This provides at least that minimal types of local typings are unique.

Lemma 5.2 (Minimal Type Uniqueness): Let $S \in \{L, H\}$. If $T; S \vdash_{\text{ML}} o : (A, \delta)$ and $T; S \vdash_{\text{ML}} o : (A', \delta')$, then $\delta = \delta'$.

A slightly stronger form of that previous lemma exists for H PCs.

Lemma 5.3 (High PC Uniqueness): If $T; H \vdash o : (A, \delta)$ and $T; H \vdash_{\text{ML}} o : (A, \delta')$, then $\delta = \delta'$.

Using slight generalization and contraposition, the previous lemma can be strengthened to the following key lemma for confinement.

Lemma 5.4 (Abstract Confinement): If $T; S \vdash o : (A, \delta)$ and $\delta_0 \sqsubset \delta$ and $T; S_0 \vdash_{\text{ML}} o : (A, \delta_0)$, then $S_0 = L$.

The following key fact, about the minimal type for futures provides the anchor to apply Abstract Confinement and arrive at Confinement.

Proposition 5.5 (Minimal Future Type): Let $\vdash C : \langle \Gamma_{act}, \Gamma_{fut}, sec \rangle$, $f_k \in \text{dom}(\Gamma_{fut})$, and $\alpha = futact(f_k)$ the home activity of f_k. Then

$$\langle \Gamma_{act}, \Gamma_{fut}, sec \rangle, A_\alpha(futlab(f_k)) \vdash_{\text{ML}} f_k : \Gamma_{act}(\alpha)\,.$$

Theorem 2 (Confinement): If a future f_k is typeable with an arbitrary PC S as of type δ strictly larger than the global level of f_k's home activity α, then f_k has been initially generated from a call to an L method of α. Formally, let $\vdash C : \langle \Gamma_{act} \cup sec, \Gamma_{fut} \rangle$, $\alpha[Q, a] \in C$, and $f_k \in \text{dom}(Q)$ with

$$\langle \Gamma_{act}, \Gamma_{fut}, sec \rangle, T; S \vdash f_k : (A_x, \delta) \text{ where } \Gamma_{act}(\alpha) \sqsubset \delta\,.$$

Then

$$A_\alpha(futlab(f_k)) = L\,.$$

The proof of confinement is basically just a combination of Lemma 5.4 and Proposition 5.5. The chain of lemmas and confinement have been proved in Isabelle/HOL as well.

C. Noninterference

Confinement can be considered as a simple security property because it is similar to a safety property: confinement is preserved on every trace of execution of a configuration. Intuitively it seems to imply confidentiality of private parts but this is only true for direct information flows. Confidentiality necessitates that no information is leaked to an outsider even considering implicit information flows as described in Section III-C. Based on those observations, we define the general property of confidentiality as *noninterference*, informally meaning that an attacker cannot learn anything despite his ability to observe configurations on all runs while comparing values that he can see: a difference in the value of the same call allows deductions about a change in hidden parts. The formal definition of noninterference for active objects in general [21] is a bisimulation over the indistinguishability relation $\sim_\alpha$ on configurations. We omit the rather technical definition of indistinguishability referring to Appendix D. Essentially, indistinguishability says that C and C_1 appear equal to the attacker α's viewpoint even if they differ in secret parts; noninterference means that this appearance is preserved by the evaluation of configurations.

Definition 5.6 (α-Noninterference): If configuration C is indistinguishable to any C_1 for α with respect to *sec* and remains so under the evaluation of configurations $\rightarrow_\parallel$, then C is α-noninterfering. Formally, we define α-noninterference C *sec* as follows.

$$\left(\begin{array}{c} C \rightarrow_\parallel C' \\ C \sim_\alpha C_1 \end{array} \right) \Longrightarrow \exists\, C_1'.\, \left(\begin{array}{c} C_1 \rightarrow_\parallel^* C_1' \\ C' \sim_\alpha C_1' \end{array} \right)$$

A main result for our security type system is *soundness*: a well-typed configuration is secure; α-noninterference holds for the configuration, i.e., it does not leak information.

Theorem 3 (Soundness): For any well-typed configuration C, we have noninterference with respect to $\alpha \in C$, i.e.,

$$\left(\begin{array}{c} \vdash C : \langle \Gamma_{act}, \Gamma_{fut}, sec \rangle \\ \alpha[Q, a] \in C \end{array} \right) \Rightarrow \alpha\text{-noninterference } C \textit{ sec}\,.$$

The proof of this theorem is a case analysis distinguishing the cases where a reduction step of the configuration has happened in the α-visible part or outside it. In the latter case, a difference in the visible part would mean a breach of confinement. Within the visible part, a straightforward case analysis shows that what is possible in one configuration must also be possible in the other, indistinguishable, one, since those parts are isomorphic; hence the same reduction rules apply. We have formalized the definitions of indistinguishability, noninterference, and multi-lateral security, as well as the statements of the theorems in Isabelle/HOL – only the soundness proof is not yet formalized but a detailed paper proof is contained in Appendix E.

The parameterization of the attacker as an active object α grants the possibility to adapt the noninterference predicate. If we universally quantify α in our definition of noninterference, we obtain a predicate where each object could be the attacker corresponding to multi-lateral security.

Definition 5.7 (Multi-Lateral Security): If a configuration C is α-noninterfering for all $\alpha \in \mathrm{dom}(C)$ then multi-lateral security holds for C.

Since no α is fixed in the type statement, the soundness theorem holds for any α if the configuration is well-typed. Hence, well-typing implies immediately multi-lateral security.

VI. Related Work and Conclusions

The main difference of our approach is that we specifically address functional active objects. We also use a non-standard security model [21] for multi-lateral security tailored to distributed active objects. Other work on actor security, e.g. [20], is based on message passing models different to our high level language model. The paper [4] addresses only direct information flows in active objects. The priority program Reliably Secure Software Systems (RS3) of the German Research Foundation (DFG) [23] addresses in its part project MoVeSPAcI [29] security of actor systems using an event based approach without futures.

The Distributed Information Flow Control (DIFC) approach [27] provides support for Java programs (Jif) to annotate programs with labels "Alice" and "Bob" for information flow control. In this approach objects are not first class citizen. The formal model [37] uses a lambda calculus λ_{DSec} to accommodate the rich hierarchy of labels but (Java) objects are not in the calculus. They use an elegant approach to prove noninterference of a type system for labels pioneered by [30] Pottier and Simonet. This approach does not apply to parallel languages since the evaluation order of parallel processes is not deterministic. The language based approach offers the first model of language based information flow control for concurrency [36], later refined by Boudol and Castellani addressing scheduling problems and related timing leaks. Many works have followed this methodology (see [35] for an overview). However, most works consider imperative while languages with various extensions like multi threading. Sabelfeld and Mantel consider message passing in distributed programs [34], [24]. These works use the secure channel abstraction, i.e. connecting remote processes of the same security class via secure channels integrating security primitives.

Distributed security has also been considered in many works in the setting of process algebras most prominently using pi calculus by [26] (see [15] and [33] providing overviews). Commonalities of process algebra based security to our work are the bisimulation notion of noninterference and asynchronous communication. There is a line of research on mobile calculi that use purely functional concurrent calculi. A few representative papers are by [19] on the pi calculus and [17] for the security pi calculus. An impressive approach on information flows for distributed languages with mobility and states [25] first introduces declassification. Similar work is by [5] also studying noninterference for distribution and mobility for Boxed Ambients. In common with these works are modeling distributed system by a calculus but none of the pi calculus related work focuses on active objects while we do not consider cryptographic primitives. An interesting perspective would be to investigate the relationship between confinement and effects of cryptographic primitives. We also use a bisimulation-based equivalence relation to express noninterference. In the applied pi calculus, for example, the notion of a static equivalence, similar to our indistinguishability is used in addition to observational equivalence that corresponds to our notion of noninterference (see e.g. [9]).

This work presented a formal framework for the security of active objects based on the semi-lattice security model that propagates confinement. We presented a safe security type system, that verifies the confinement property and is sound, i.e., checks security, with respect to a dedicated formal notion of noninterference, or more generally, multi-lateral security. ASP_{fun} makes secure down-calls possible and is still applicable bi-directionally as illustrated by implementing the NSPK protocol. The proofs have been in large parts formalized in Isabelle/HOL. An implementation of functional active objects is given by Erlang Active Objects in [13]. providing a simple extension by a run-time monitor for confinement. [14].

REFERENCES

[1] M. Abadi and L. Cardelli, *A Theory of Objects*. Springer-Verlag, New York, 1996.

[2] G. Agha, I. A. Mason, S. F. Smith, and C. L. Talcott, "Towards a theory of actor computation (extended abstract)," in *CONCUR'92: Proceedings of the Third International Conference on Concurrency Theory*, W. R. Cleaveland, Ed. Berlin, Heidelberg: Springer-Verlag, 1992, pp. 565–579.

[3] R. Anderson, *Security Engineering – A Guide to Building Dependable Distributed Systems*. Wiley, 2001.

[4] I. Attali, D. Caromel, L. Henrio, and F. L. D. Aguila, "Secured information flow for asynchronous sequential processes," *Electr. Notes Theor. Comput. Sci.*, vol. 180, no. 1, pp. 17–34, 2007.

[5] M. Bugliesi, G. Castagna, and S. Crafa, "Access control for mobile ambients: the calculus of boxed ambients," *ACM Transactions on Programming Languages and Systems*, vol. 26, no. 1, pp. 57–124, 2004.

[6] D. Caromel, C. Delbé, A. di Costanzo, and M. Leyton, "ProActive: an integrated platform for programming and running applications on grids and P2P systems," *Computational Methods in Science and Technology*, vol. 12, no. 1, pp. 69–77, 2006.

[7] D. Caromel, L. Henrio, and B. P. Serpette, "Asynchronous and deterministic objects," in *Proceedings of the 31st ACM SIGACT-SIGPLAN symposium on Principles of programming languages*. ACM Press, 2004, pp. 123–134.

[8] B. A. Davey and H. A. Priestley, *Introduction to Lattices and Order (2. ed.)*. Cambridge University Press, 2002.

[9] S. Delaune, S. Kremer, and M. D. Ryan, "Symbolic bisimulation for the applied pi calculus," *Journal of Computer Security*, vol. 18, no. 2, pp. 317–377, 2010.

[10] D. E. Denning, "Lattice model of secure information flow," *Communications of the ACM*, vol. 19, no. 5, pp. 236–242, 1976.

[11] D. E. Denning and P. J. Denning, "Certification of programs for secure information flow," *Communications of the ACM*, vol. 20, no. 7, 1977.

[12] J. S. Fenton, "Information protection systems," Ph.D. dissertation, University of Cambridge, 1973.

[13] A. Fleck and F. Kammüller, "Implementing privacy with erlang active objects," in *5th International Conference on Internet Monitoring and Protection, ICIMP'10*. IEEE, 2010.

[14] ——, "A security model for functional active objects with an implementation in erlang," in *Computational Informatics, Social Factors and New Information Technologies: Hypermedia Perspectives and Avant-Garde Experiences in the Era of Communicability Expansion*. Blue Herons, 2011.

[15] R. Focardi and R. Gorrieri, "A taxonomy of security properties for process algebras," *Journal of Computer Security*, vol. 3, no. 1, pp. 5–34, 1995.

[16] R. H. Halstead, Jr., "Multilisp: A language for concurrent symbolic computation," *ACM Transactions on Programming Languages and Systems (TOPLAS)*, vol. 7, no. 4, pp. 501–538, 1985.

[17] M. Hennessy and J. Riely, "Information flow vs. resource access in the asynchronous pi-calculus," *ACM Trans. Program. Lang. Syst.*, vol. 24, no. 5, pp. 566–591, 2002.

[18] L. Henrio, F. Kammüller, and B. Lutz, "Aspfun: A typed functional active object calculus," *Science of Computer Programming*, vol. 77, no. 7-8, pp. 823–847, 2012.

[19] K. Honda, V. T. Vasconcelos, and N. Yoshida, "Secure information flow as typed process behaviour," in *ESOP*, ser. Lecture Notes in Computer Science, G. Smolka, Ed., vol. 1782. Springer, 2000, pp. 180–199.

[20] D. Hutter, H. Mantel, I. Schaefer, and A. Schairer, "Security of multi-agent systems: A case study on comparison shopping," *J. Applied Logic*, vol. 5, no. 2, pp. 303–332, 2007.

[21] F. Kammüller, "A semi-lattice model for multi-lateral security," in *Data Privacy Management, DPM'12, Seventh International Workshop. Co-located with ESORICS'12*, ser. LNCS, Security and Cryptology, vol. 7731. Springer, 2013.

[22] G. Lowe, "An attack on the needham-schroeder public-key authentication protocol," *INFORMATION PROCESSING LETTERS*, vol. 56, pp. 131–133, 1995.

[23] H. Mantel, D. Hutter, G. Snelting, and T. Nipkow, "Reliably secure software systems – Priority Program of the German Research Foundation (DFG)," 2010, http://www.reliably-secure-software-systems.de.

[24] H. Mantel and A. Sabelfeld, "A unifying approach to the security of distributed and multi-threaded programs," *J. Computer Security*, vol. 11, p. 2003, 2002.

[25] A. A. Matos and J. Cederquist, "Non-disclosure for distributed mobile code," *Mathematical Structures in Computer Science*, vol. 21, no. 6, pp. 1111–1181, 2011.

[26] R. Milner, *Communication and concurrency*, ser. PHI Series in computer science. Prentice Hall, 1989.

[27] A. C. Myers and B. Liskov, "A decentralized model for information flow control," in *Proceedings of the sixteenth ACM symposium on Operating systems principles*, ser. SOSP '97. New York, NY, USA: ACM, 1997, pp. 129–142. [Online]. Available: http://doi.acm.org/10.1145/268998.266669

[28] R. M. Needham and M. D. Schroeder, "Using encryption for authentication in large networks of computers," *Communications of the ACM*, no. 21, 1978.

[29] A. Poetzsch-Heffter and C. Feller, "Modular verification of security properties in actor implementations (movespaci)," 2011, project of the German Research Foundation, Priority Program RS3. [Online]. Available: http://softech.informatik.uni-kl.de/Homepage/MoVeSPAcI

[30] F. Pottier and V. Simonet, "Information flow inference for ml," *ACM Trans. Program. Lang. Syst.*, vol. 25, no. 1, pp. 117–158, 2003.

[31] C. W. Probst and R. R. Hansen, "An extensible analysable system model," *Information Security Technical Report*, vol. 13, no. 4, pp. 235–246, Nov. 2008.

[32] C. W. Probst, J. Hunker, D. Gollmann, and M. Bishop, Eds., *Insider Threats in Cybersecurity*. Springer, 2010.

[33] P. Ryan, S. Schneider, M. Goldsmith, G. Lowe, and B. Roscoe, *The Modelling and Analysis of Security Protocols*. Addison-Wesley, 2000.

[34] A. Sabelfeld and H. Mantel, "Static confidentiality enforcement for distributed programs." in *Static Analysis Symposium, SAS'02*, ser. LNCS, vol. 2477. Springer, 2002.

[35] A. Sabelfeld and A. C. Myers, "Language-based information-flow security," *Selected Areas in Communications*, vol. 21, pp. 5–19, 2003.

[36] G. Smith and D. Volpano, "Secure information flow in a multi-threaded imperative language," in *POPL'98*. ACM, 1998.

[37] L. Zheng and A. C. Myers, "Dynamic security labels and static information flow control," *International Journal of Information Security*, vol. 6, no. 2–3, 2007.

APPENDIX

A: Booleans and conditional in the ς-calculus and their security types

To prepare for the type inference, we need the implementation of the boolean datatype and the *if-then-else* in the ς-calculus, i.e. in the

local calculus of ASP_{fun}.

$$\begin{aligned}
\text{true} &= [\,\text{if} = \varsigma(y)\mathit{this}.\text{then}(y),\\
&\qquad \text{then} = \varsigma(y)[], \text{else} = \varsigma(y)[]\,]\\
\text{false} &= [\,\text{if} = \varsigma(y)\mathit{this}.\text{else}(y),\\
&\qquad \text{then} = \varsigma(y)[], \text{else} = \varsigma(y)[]\,]\\
\text{if } b \text{ then } c \text{ else } d &=\\
&\qquad ((b.\text{then} := \varsigma(y)c).\text{else} := \varsigma(y)d).\text{if}([])
\end{aligned}$$

In the third line above, $\mathit{this}, y \notin FV(c) \cup FV(d)$; [] denotes the empty object. The definition shows how – similar to λ-calculus – the functionality of the constructor is encoded in the elements of the datatype: when b is true its method if delegates to the method then, filled with term c, when false, if delegates to else, executing term d. Typing of the *if-then-else* construct is a base test for an information flow type system as this construct is the basic example that gives rise to implicit information flows. We will thus here illustrate how the security type rules presented in this paper establish that the guard of the *if-then-else* construct, the if, must be typed with the same PC as the branches, i.e., then and else. Then it immediately follows that if the method if has H-PC then the branches must have H-PC as well. The reasoning instantiates type rules showing the constraints that follow for the security assignment in the security type Σ_{ifte}. A condition b in the method if of an *if-then-else* object evaluates to either $true$ or $false$. We consider those two possibilities and infer their types and the resulting constraints. To type *true*, we initially type *this* which can be done only by rule VAL SELF leading to the following typing where $\Sigma_{\text{ifte}} = (A_{\text{ifte}}, \delta_{\text{ifte}})$ is the security type for the *if-then-else* object and $M_{\text{ifte}} = \sqcup\{A_{\text{ifte}}(\text{if}), A_{\text{ifte}}(\text{then}), A_{\text{ifte}}(\text{else})\}$. We use the arbitrary set of additional type assumptions T provided by the rule to integrate the type assumption for y already here. It is needed further down for typing the object but only formally.

INSTANCE VAL SELF
$$\mathit{this} : \Sigma_{\text{ifte}} :: (y : \Sigma_{\text{ifte}}) :: T; M_{\text{ifte}} \vdash \mathit{this} : \Sigma_{\text{ifte}}$$

We then apply the rule TYPE CALL to infer a type for *this*.then. The following instance of that rule sets the parameters such that it can be applied to the previous INSTANCE VAL SELF.

INSTANCE TYPE CALL
$$\frac{\begin{array}{c}\mathit{this} : \Sigma_{\text{ifte}} :: y : \Sigma_{\text{ifte}} :: T; M_{\text{ifte}} \vdash \mathit{this} : \Sigma_{\text{ifte}}\\ \mathit{this} : \Sigma_{\text{ifte}} :: y : \Sigma_{\text{ifte}} :: T; A_{\text{ifte}}(\text{then}) \vdash [] : \Sigma_{\text{ifte}}\end{array}}{\mathit{this} : \Sigma_{\text{ifte}} :: y : \Sigma_{\text{ifte}} :: T; A_{\text{ifte}}(\text{then}) \vdash \mathit{this}.\text{then}([]) : \Sigma_{\text{ifte}}}$$

Now, to type the *true* object including its fields then and else we next need an instance of TYPE OBJECT.

INSTANCE TYPE OBJECT
$$\frac{\begin{array}{c}\mathit{this} : \Sigma_{\text{ifte}} :: y : \Sigma_{\text{ifte}} :: T; A_{\text{ifte}}(\text{if}) \vdash \mathit{this}.\text{then}([]) : \Sigma_{\text{ifte}}\\ \mathit{this} : \Sigma_{\text{ifte}} :: y : \Sigma_{\text{ifte}} :: T; A_{\text{ifte}}(\text{then}) \vdash [] : \Sigma_{\text{ifte}}\\ \mathit{this} : \Sigma_{\text{ifte}} :: y : \Sigma_{\text{ifte}} :: T; A_{\text{ifte}}(\text{else}) \vdash [] : \Sigma_{\text{ifte}}\end{array}}{T; M_{\text{ifte}} \vdash \text{true} = [\,\text{if} = \varsigma(y)\mathit{this}.\text{then}(y), \text{then} = \varsigma(y)[], \text{else} = \varsigma(y)[]\,] : \Sigma}$$

The main observation is that the following constraint must hold for A_{ifte}

$$A_{\text{ifte}}(\text{if}) = A_{\text{ifte}}(\text{then})$$

because this is necessary for the first proviso of the above instance to be matched with the previous type derivation for *this*.then([]) by INSTANCE TYPE CALL.

With a very similar argument for typing *false*, i.e., *this*.else([]), we arrive at a similar constraint.

$$A_{\text{ifte}}(\text{if}) = A_{\text{ifte}}(\text{else})$$

Since for an arbitrary *if-then-else* guard b we have to allow both values *true* and *false* as possible outcome we have to combine the constraints and conclude for A_{ifte} the following overall constraint.

$$A_{\text{ifte}}(\text{if}) = A_{\text{ifte}}(\text{then}) = A_{\text{ifte}}(\text{else})$$

The update of the methods then and else does not change the PC and thus preserves the security assignment and the constraints. This constraint is what we expect for information flow security. If the guard of an *if-then-else* can only be typed in a H-PC then its branches must also be "lifted" to H. Only if the guard can be typed in a L-PC, can the branches also be typed in L-PC.

Note on typing constants

In the above type rule instances we have used typings for constants, for example, the empty object [] as granted and did not refine them any further.

$$\mathit{this} : \Sigma_{\text{ifte}} :: y : \Sigma_{\text{ifte}} :: T; A_{\text{ifte}}(\text{else}) \vdash [] : \Sigma_{\text{ifte}}$$

A word is in order to explain how these are constructed and their types are derived. A simple way to integrate the empty object into an activity is to add a method empty and then replace all [] by *this*.empty. The security assignment should be $A(\text{empty}) = L$. Then, we can use TYPE CALL to have $\ldots L \vdash \mathit{this}.\text{empty} : \Sigma_{\text{ifte}}$ and from there derive the above $\ldots; A_{\text{ifte}}(\text{else}) \vdash \mathit{this}.\text{empty} : \Sigma_{\text{ifte}}$. However, [] (and other commonly used plain objects) can more practically be considered as *activities without any H methods* that are included as a "data base" in a configuration. Then, an occurrence of [] is literally the activity named "empty object", i.e., [] is an activity reference. For the typing, the natural type of the empty object is given as the empty security assignment $\varnothing$, i.e., the partial function that is undefined for all inputs, and the bottom element $\bot$ of the visibility semi-lattice which corresponds to the empty set of activity names.

$$\langle \Gamma_{\mathit{act}}, \Gamma_{\mathit{fut}}, \mathit{sec} \rangle; L \vdash [] : (\varnothing, \bot)$$

By definition this typing with L-PC as $(\varnothing, \bot)$ for the empty object enables typing [] "into" any other activity type (A, δ) because $\varnothing \subseteq A$ and $\bot \sqsubseteq \delta$. Thus – by SECASS SUBSUMPTION and GLOB SUBSUMPTION – $\ldots; L \vdash [] : (A, \delta)$.

A similar type and subtyping argumentation goes for other constants, for example 0 or 1, used in the running example. Similar to Church numerals simple term representation can be given to them in ASP_{fun}. Such constant activities η must have their methods all assigned to L, i.e., their security assignment A maps all method names of η to L. Then the PC of the activity η is also L because it is given as $\sqcup\{L\}$ according to the rule TYPE CONFIGURATION. The global level of a constant activity like η is defined as the set $\{eta\}$. If the constant η is used by referencing it in other activities of the configuration, the name η becomes part of the other activities' global levels.

B: Running Example – Details on Typing

The following shows why the example configuration presented as running example cannot be typed with income $\mapsto H \in \mathit{sec}$.

Implementation: The quicksort function is described in Section II-C. The manager activity that controls the ordering of a list and the sorting object χ that calls the ord method in β-objects are repeated here for convenience of the reader.

$$\begin{aligned}
\alpha\big[\varnothing, [&\text{manage} = \varsigma(y)\mathit{this}.\text{sort}(\mathit{this}.\text{list}),\\
&\text{sort} = \varsigma(y)\ \chi.\text{qsort}(y),\\
&\text{list} = \ldots]\big]
\end{aligned}$$

$$\chi[\varnothing,$$
$$\begin{array}{l}[\text{qsort} = \varsigma(y)\ \text{if}\ y = [\,]\ \text{then}\ [\,]\\ \qquad\qquad \text{else let}\ (a :: l) = y\\ \qquad\qquad\qquad (l_1, l_2) = \mathit{this}.\text{part}\ (a, l)\\ \qquad\qquad\qquad l_1' = \text{if}\ \#l_1 \leq 1\ \text{then}\ l_1\ \text{else}\ \mathit{this}.\text{qsort}(l_1)\\ \qquad\qquad\qquad l_2' = \text{if}\ \#l_2 \leq 1\ \text{then}\ l_2\ \text{else}\ \mathit{this}.\text{qsort}(l_2)\\ \qquad\qquad \text{in}\ l_1' @ [a] @ l_2'\\ \qquad\qquad \text{end},\\ \text{part} = \varsigma(p, y)\ \text{if}\ y = [\,]\ \text{then}\ ([\,], [\,])\\ \qquad\qquad \text{else let}\ (a :: l) = y\\ \qquad\qquad\qquad (l_1, l_2) = \mathit{this}.\text{part}\ (p, l)\\ \qquad\qquad\quad \text{in if}\ p < a.\text{ord then}\ (l_1, a :: l_2)\ \text{else}\ (a :: l_1, l_2)\\ \qquad\qquad \text{end}\\ ]]\end{array}$$

The extended method ord that bears a dependency between ord and income,

$$\beta[\varnothing, \quad [\text{ord} = \varsigma(y)\ \text{if}\ \mathit{this}.\text{income}/10^3 \geq 1\ \text{then}\ 1\ \text{else}\ 0],\ \text{income} = \ldots \quad]$$

is not typeable for any security assignment *sec* that imposes the constraint that method income $\mapsto H$. The following type inference elaborates that the type system rejects any security assignment that contains the constraint income $\mapsto H$. It illustrates how the security assignment $A_{\beta_i} = \mathit{ass}(\Gamma_{\mathit{act}}(\beta_i))$ is inferred.

Typing remote call implies ord $\mapsto L$

Since the method ord is called remotely in χ via α we need that ord $\mapsto L$, which cannot be possible because of the dependency in the above implementation. To be able to type the call to β_i.ord in the object χ this method must be an L-method according to TYPE CALL and GLOB SUBSUMPTION. More precisely, let $(A_{\beta_i}, \delta_{\beta_i}) = \Gamma_{\mathit{act}}(\beta_i)$. We have that $\langle \Gamma_{\mathit{act}}, \Gamma_{\mathit{fut}}, \mathit{sec}\rangle, \varnothing; M_{\beta_i} \vdash \beta_i : (A_{\beta_i}, \delta_{\beta_i})$ because of TYPE ACTIVE OBJECT REFERENCE and $\beta_i \in \text{dom}(C)$. The PC is $M_{\beta_i} = \sqcup_{j \in \text{dom} A_{\beta_i}} A_{\beta_i}(j)$ where A_{β_i} needs to be inferred in the process. We can use next TYPE CALL to type $\langle \Gamma_{\mathit{act}}, \Gamma_{\mathit{fut}}, \mathit{sec}\rangle, \varnothing; A_{\beta_i}(\text{ord}) \vdash \beta_i.\text{ord} : (A_{\beta_i}, \delta_{\beta_i})$. However, to type β_i.ord in the context of the object χ it needs to be typed as $(A_{\beta_i}, \delta_\chi)$ with global type component δ_χ. This upgrading of the call can only be achieved by application of rule GLOB SUBSUMPTION which requires that $\delta_{\beta_i} \sqsubseteq \delta_\chi$ which is true but also requires that the PC of the typing $\langle \Gamma_{\mathit{act}}, \Gamma_{\mathit{fut}}, \mathit{sec}\rangle, \varnothing; A_{\beta_i}(\text{ord}) \vdash \beta_i.\text{ord} : (A_{\beta_i}, \delta_{\beta_i})$, i.e., $A_{\beta_i}(\text{ord})$, is L.

Typing β_i.ord at global level δ_{β_i} only with H-PC

The next part of the argument states that the only type that can be inferred for a call β_i.ord is $T; H \vdash \beta_i.\text{ord} : (A_{\beta_i}, \delta_{\beta_i})$, i.e., with H-PC. This is because types for calls can only be inferred by rule CALL and $A_{\beta_i}(\text{ord}) = H$ which coerces the PC according to rule CALL to H. For clarity of the exposition, we omit in the following $\langle \Gamma_{\mathit{act}}, \Gamma_{\mathit{fut}}, \mathit{sec}\rangle$ in front of the typings. Since we want to arrive at $\vdash C : \langle \Gamma_{\mathit{act}}, \Gamma_{\mathit{fut}}, \mathit{sec}\rangle$, by the inversion principle of inductive type definitions, all provisos of TYPE CONFIGURATION have to be true. Since

$$\beta_i[\varnothing, [\text{ord} = \varsigma(y)\ t_{\text{ord}}, \text{income} = \varsigma(y)\ t_{\text{income}}]] \in C,$$

the fourth proviso, first clause, of TYPE CONFIGURATION yields

$$\varnothing, M_{\beta_i} \vdash [\text{ord} = \varsigma(y)\ t_{\text{ord}}, \text{income} = \varsigma(y)\ t_{\text{income}}] : (A_{\beta_i}, \delta_{\beta_i}).$$

The coercion of $A_{\beta_i}(\text{ord})$ to H is a consequence of the typing of the object β_i with an instance of rule TYPE OBJECT.

INSTANCE TYPE OBJECT

$$\frac{\begin{array}{c}\mathit{this} : \Sigma_{\beta_i} :: [\,] : \Sigma_{\beta_i} :: \varnothing; A_{\beta_i}(\text{ord}) \vdash t_{\text{ord}} : \Sigma_{\beta_i}\\ \mathit{this} : \Sigma_{\beta_i} :: [\,] : \Sigma_{\beta_i} :: \varnothing; A_{\beta_i}(\text{income}) \vdash t_{\text{income}} : \Sigma_{\beta_i}\end{array}}{\varnothing; M_{\beta_i} \vdash [\text{ord} = \varsigma(y)\ t_{\text{ord}}, \text{income} = \varsigma(y)\ t_{\text{income}}] : \Sigma_{\beta_i}}$$

This instance enforces $A(\text{ord})$ to be the same as the PC in the typing of

$$\begin{array}{rcl} t_{\text{ord}} & = & \text{if}\ \mathit{this}.\text{income}/10^3 \geq 1\ \text{then}\ 1\ \text{else}\ 0\\ & = & (((\mathit{true}.\text{if} := (\mathit{this}. > 0(\mathit{this}.\text{div}_{10^3}(\mathit{this}.\text{income})))\\ & &).\text{then} := 1).\text{else} := 0).\text{if}([\,]).\end{array}$$

The only way to arrive at a type for t_{ord} is by an application of TYPE CALL as in the following instance.

INSTANCE TYPE CALL

$$\frac{\begin{array}{c}T; S \vdash (((\mathit{true}.\text{if} := (\mathit{this}. > 0(\mathit{this}.\text{div}_{10^3}(\mathit{this}.\text{income}))))\\ .\text{then} := 1).\text{else} := 0) : \Sigma\\ T; A_{\beta_i}(\text{if}) \vdash [\,] : \Sigma\end{array}}{T; A_{\beta_i}(\text{if}) \vdash t_{\text{ord}} : \Sigma}$$

In order to match the conclusion of the above with the first proviso of the earlier INSTANCE TYPE OBJECT, the security assignment of ord is coerced to that of if

$$A_{\beta_i}(\text{ord}) = A_{\beta_i}(\text{if}).$$

We only need to show that $A_{\beta_i}(\text{if}) = H$ and we are finished.

Typing implies that $A_{\beta_i}(\mathit{if}) = H$

The following chain of steps shows how a type for the body of ord and thus $A_{\beta_i}(\text{ord})$ must be inferred detailing how the security assignment parameter $A_{\beta_i}(\text{if})$ needs to be instantiated to H. The chain of reasoning starts from the one specified security assignment income $\mapsto H$ in *sec* and shows that then also ord $\mapsto H$ which contradicts the above ord $\mapsto L$. Hence, no type can exist with the constraint income $\mapsto H$ for this configuration.

$A_{\beta_i}(\text{income})$ is H by constraint on sec and thus A_{β_i}. According to VAL SELF with

$$M = \sqcup\{A_{\beta_i}(\text{income}), A_{\beta_i}(\text{ord}), \ldots\} = H$$

we get the following typing for *this*.

$$T; H \vdash \mathit{this} : (A_{\beta_i}, \delta_{\beta_i})$$

According to TYPE CALL, *this*.income is typeable only with PC as H since $\{\text{income} \mapsto H\} \subseteq A_{\beta_i}$.

$$T; H \vdash \mathit{this}.\text{income} : (A_{\beta_i}, \delta_{\beta_i})$$

The previous typing feeds into rule TYPE CALL again but this time for the parameter t. Since the PC S_j matches with H we get again a H-PC coercing the method div_{10^3} also to be assigned to H in A_{β_i}.

$$T; H \vdash \mathit{this}.\text{div}_{10^3}(\mathit{this}.\text{income}) : (A_{\beta_i}, \delta_{\beta_i})$$

In the same fashion, the previous considered as a parameter typing TYPE CALL consequently coerces $A_{\beta_i}(> 0)$ also to H:

$$T; H \vdash \mathit{this}. > 0(\mathit{this}.\text{div}_{10^3}(\mathit{this}.\text{income})) : (A_{\beta_i}, \delta_{\beta_i})$$

We instantiate UPDATE as follows.

INSTANCE TYPE UPDATE

$$\frac{\begin{array}{c}\varnothing; S \vdash \text{true} : (A_{\beta_i}, \delta_{\beta_i})\\ \mathit{this} : \Sigma_{\beta_i} :: [\,] : \Sigma_{\beta_i} :: \varnothing; A_{\beta_i}(\text{if}) \vdash\\ \mathit{this}. > 0(\mathit{this}.\text{div}_{10^3}(\mathit{this}.\text{income})) : (A_{\beta_i}, \delta_{\beta_i})\end{array}}{\begin{array}{c}\varnothing; \sqcup\{A_{\beta_i}(\text{if}), \ldots\} \vdash\\ \text{true}.\text{if}[\,] := \mathit{this}. > 0(\mathit{this}.\text{div}_{10^3}(\mathit{this}.\text{income})) : (A_{\beta_i}, \delta_{\beta_i})\end{array}}$$

The first proviso, the typing for *true* can be inferred as shown in the previous section, using rule SEC ASS SUBSUMPTION in addition to embed it into β_i. We spell out some portion of $M_{\beta_i} = \sqcup\{A_{\beta_i}(\text{if}), \ldots\}$ above to emphasize that $A_{\beta_i}(\text{if})$ is part of the PC. The dots stand for $A_{\beta_i}(\text{ord})$ and $A_{\beta_i}(\text{income})$ etc. To match the previous derivation above of $\varnothing; H \vdash \mathit{this}. > 0(\mathit{this}.\text{div}_{10^3}(\mathit{this}.\text{income}))$:

$(A_{\beta_i}, \delta_{\beta_i})$ to the second proviso in the above instance it is necessary to coerce

$$A_{\beta_i}(\text{if}) = H.$$

We are finished here already because we have already shown above that $A_{\beta_i}(\text{if}) = A_{\beta_i}(\text{ord})$ which thus is H contradicting the earlier requirement to be L.

For completeness, we continue the derivation of the body of t_{ord}. From the previous step above, we get the conclusion

$$\varnothing; H \vdash \text{true.if}[] := \mathit{this}.\ > 0(\mathit{this}.\text{div}_{10^3}(\mathit{this}.\text{income})) : (A_{\beta_i}, \delta_{\beta_i}).$$

We can again instantiate the update rule.

INSTANCE TYPE UPDATE

$$\frac{\varnothing; H \vdash \text{true.if}[] := \mathit{this}.\ > 0(\mathit{this}.\text{div}_{10^3}(\mathit{this}.\text{income})) : (A_{\beta_i}, \delta_{\beta_i}) \quad \mathit{this} : (A_{\beta_i}, \delta_{\beta_i}) :: y : (A_{\beta_i}, \delta_{\beta_i}) :: \varnothing; H \vdash 1 : (A_{\beta_i}, \delta_{\beta_i})}{\varnothing; H \vdash \mathit{this}.\ > 0(\mathit{this}.\text{div}_{10^3}(\mathit{this}.\text{income})).\text{then} := 1 : (A_{\beta_i}, \delta_{\beta_i})}$$

And a second time we instantiate TYPE UPDATE for 0 to finally obtain

$$\varnothing; H \vdash ((\mathit{this}.\ > 0(\mathit{this}.\text{div}_{10^3}(\mathit{this}.\text{income}))).\text{then} := 1).\text{else} := 0 : (A_{\beta_i}, \delta_{\beta_i})$$

Running Example: Typing Summary

The coercions revealed in the above steps determine the parameter A_{β_i} in summary as follows.

$$A_{\beta_i} = [\text{income} \mapsto H,\ \text{div}_{10^3} \mapsto H,\ > 0 \mapsto H\ \text{if} \mapsto H,\ \text{ord} \mapsto S_{\text{if}}]$$

I.e., the only possible instantiation for $A_{\beta_i}(\text{ord})$ is H. We cannot meet the required constraint $A_{\beta_i}(\text{ord}) = L$ necessary to call it from the outside in χ as explained initially. Therefore, the example configuration cannot be typed with the constraint income $\mapsto H$.

Borderline Example for Confinement

The confinement property states that remote calls can only be addressed to L methods. But does this simple security property guarantee that no hidden H methods can be returned with the reply to such a call? Consider the following example

$$\alpha[\varnothing, [\text{leak} = \varsigma(y)\mathit{this}, \text{key} = \varsigma(y)n]]$$

where n is an integer representing a secret key. Let the security assignment for α be $\{\text{leak} \mapsto L, \text{key} \mapsto H\}$. One may think that an activity β could contain a call α.leak.key since the method leak is L enabling the remote call to α.leak. Once the call result is returned into β, it would evaluate to the active object of α inside β (since *this* represents this active object of α). Since we are now already in β, it might seem possible to apply the method key to extract the key.

How does the security type system prevent this? Since the typing for *this* inside α is only possible with the PC as $M_\alpha = \sqcup_{i \in \{\text{leak, key}\}} A_\alpha(i) = H$ (since $A_\alpha(\text{key}) = H$), the typing for *this* yields

INSTANCE VAL SELF $\quad \mathit{this} : \Sigma_\alpha :: T; H \vdash \mathit{this} : (A_\alpha, \delta_\alpha).$

Typing the object α must use the following instance of TYPE OBJECT.

INSTANCE TYPE OBJECT

$$\frac{\mathit{this} : \Sigma_\alpha :: y : \Sigma_\alpha :: \varnothing; A_\alpha(\text{leak}) \vdash \mathit{this} : \Sigma_\alpha \quad \mathit{this} : \Sigma_\alpha :: y : \Sigma_\alpha :: \varnothing; A_\alpha(\text{key}) \vdash n : \Sigma_\alpha}{\varnothing; \sqcup\{A_\alpha(\text{leak}), A_\alpha(\text{key})\} \vdash [\text{leak} = \varsigma(y)\mathit{this}, \text{key} = \varsigma(y)n] : \Sigma_\alpha}$$

Now, matching the instance of VAL SELF for *this* with the first proviso of the instance of TYPE OBJECT coerces $A_\alpha(\text{leak})$ to H contradicting the initial specification. I.e., the method leak is forced to be H and cannot be called remotely.

C: Formal Semantics of ASP$_{fun}$

For a concise representation of the operational semantics, we define contexts as expressions with a single hole ($\bullet$). A context $E[t]$ denotes the term obtained by replacing the single hole by t.

$$E ::= \ \bullet \mid [l_i = \varsigma(y)E, l_j = \varsigma(y_j)t_j^{j \in (1..n)-\{i\}}] \mid E.l_i(t) \mid s.l_i(E) \mid E.l_i := \varsigma(y)s \mid s.l_i := \varsigma(y)E \mid \mathit{Active}(E)$$

This notion of context is used in the formal semantics of ASP$_{\text{fun}}$ in Table V and also in the definition of visibility (see Definition 3.1).

D: Indistinguishability

In ASP$_{\text{fun}}$ active objects are created by activation, futures by method calls. Names of active objects and futures may differ in evaluations of the same configuration but this does not convey any information to the attacker. We use "partial bijections" to express the equality of the visible parts of a configuration.

Definition 6.1 (Typed Bijection): A typed bijection is a finite partial function σ on activities α (or futures f_k respectively) such that for a type T

$$\forall\, a : \text{dom}(\sigma).\ \vdash a : T \implies \vdash \sigma(a) : T.$$

By $t[\sigma, \tau] =|_{\mathit{sec}}\ t'$ we denote the equality of terms up to replacing all occurrences of activity names α or futures f_k by their counterparts $\tau(\alpha)$ or $\sigma(f_k)$, respectively, restricted to the label names in *sec*, i.e., in the object terms t and t' we exempt those parts of the objects that are private. The local reduction with $\rightarrow_\varsigma^*$ of a term t to a value t_e (again up to future and activity references) is written as $t \Downarrow t_e$.

Definition 6.2 (Equality up to Name Isomorphism): An equality up to name isomorphism is a family of equivalence relations on ASP$_{\text{fun}}$ terms indexed by two typed bijections $(\sigma, \tau) := R$ and security assignment *sec* consisting of the following differently typed sub-relations; the sub-relation's types are indicated by the naming convention: t for ς-terms, α, β for active objects, f_k, f_j for futures, Q_α, Q_β for request queues.

$$\begin{aligned}
t =_R t' \quad &\equiv \quad t \Downarrow t_e \wedge t' \Downarrow t'_e \wedge\ t_e[\sigma, \tau] =|_{\mathit{sec}}\ t'_e \\
\alpha =_R \beta \quad &\equiv \quad \sigma(\alpha) = \beta \\
f_k =_R f_j \quad &\equiv \quad \tau(f_k) = f_j \\
Q_\alpha =_R Q_\beta \quad &\equiv \quad \left(\begin{array}{l} \text{dom}(\tau) \supseteq \text{dom}(Q_\alpha) \\ \text{ran}\ (\tau) \supseteq \text{dom}(Q_\beta) \\ \forall\, f_k \in \text{dom}(Q_\alpha). \\ Q_\alpha(f_k) =_R Q_\beta(\tau(f_k)) \end{array} \right) \\
\alpha[Q_\alpha, t_\beta] =_R \beta[Q_\beta, t_\alpha] \quad &\equiv \quad \alpha =_R \beta \wedge Q_\alpha =_R Q_\beta \wedge t_\alpha =_R t_\beta
\end{aligned}$$

Such an equivalence relation defined by two typed bijections σ and τ may exist between given sets V_0, V_1 of active object names in C, C_1. If V, V_1 correspond to the viewpoints of attacker α in C and its counterpart in C_1 we call this equivalence relation indistinguishability. In the following, we use the *visibility range* based on Definition 3.1 as $\mathit{VIsec}(\alpha, C) \equiv \{\beta \in \text{dom}(C) \mid \beta \sqsubseteq_C^{\mathit{sec}} \alpha\}$.

Definition 6.3 (Indistinguishability): Let C, C_1 be arbitrary configurations, well-typed with respect to a security specification *sec*, active object $\alpha \in \text{dom}(C)$ and $\alpha \in \text{dom}(C_1)$ (we exempt α from renaming for simplicity). Configurations C and C_1 are called indistinguishable with respect to α and *sec*, if α's visibility ranges are the same in both up to name isomorphism.

$$C \sim_\alpha C_1 \equiv \exists\, \sigma, \tau. \left(\begin{array}{l} \mathit{VIsec}(\alpha, C) = \text{dom}(\sigma) \\ \mathit{VIsec}(\alpha, C_1) = \text{ran}\ (\sigma) \\ \forall\, \beta \in \mathit{VIsec}(\alpha, C). \\ C(\beta) =_{\sigma,\tau} C_1(\sigma(\beta)) \end{array} \right)$$

As an example for α-indistinguishable configurations consider the running example. In the original (non-fallacious) form, β_1.income could be 42 in configuration C and it could be 1042 in configuration

$$\text{CALL}\quad \frac{l_i \in \{l_j\}^{j\in 1..n}}{E\left[[l_j = \varsigma(y_j)s_j]^{j\in 1..n}.l_i(t)\right] \to_\varsigma E\left[s_i\{\mathit{this} \leftarrow [l_j = \varsigma(y_j)s_j]^{j\in 1..n}, y_i \leftarrow t\}\right]}$$

$$\text{UPDATE}\quad \frac{l_i \in \{l_j\}^{j\in 1..n}}{E\left[[l_j = \varsigma(y_j)s_j]^{j\in 1..n}.l_i := \varsigma(y)t\right] \to_\varsigma E\left[[l_i = \varsigma(y)t, l_j = \varsigma(y_j)s_j^{j\in(1..n)-\{i\}}]\right]}$$

$$\text{LOCAL}\quad \frac{s \to_\varsigma s'}{\alpha[f_i \mapsto s :: Q, t] :: C \to_\parallel \alpha[f_i \mapsto s' :: Q, t] :: C}$$

$$\text{ACTIVE}\quad \frac{\gamma \notin (\mathrm{dom}(C) \cup \{\alpha\}) \qquad \mathit{noFV}(s)}{\alpha[f_i \mapsto E[\mathit{Active}(s)] :: Q, t] :: C \to_\parallel \alpha[f_i \mapsto E[\gamma] :: Q, t] :: \gamma[\varnothing, s] :: C}$$

$$\text{REQUEST}\quad \frac{f_k \text{ fresh} \qquad \mathit{noFV}(s) \qquad \alpha \neq \beta}{\alpha\left[f_i \mapsto E[\beta.l(s)] :: Q, t\right] :: \beta[R, t'] :: C \to_\parallel \alpha\left[f_i \mapsto E[f_k] :: Q, t\right] :: \beta\left[f_k \mapsto t'.l(s) :: R, t'\right] :: C}$$

$$\text{SELF-REQUEST}\quad \frac{f_k \text{ fresh} \qquad \mathit{noFV}(s)}{\alpha\left[f_i \mapsto E[\alpha.l(s)] :: Q, t\right] :: C \to_\parallel \alpha\left[f_k \mapsto t.l(s) :: f_i \mapsto E[f_k] :: Q, t\right] :: C}$$

$$\text{REPLY}\quad \frac{\beta[f_k \mapsto s :: R, t'] \in \alpha[f_i \mapsto E[f_k] :: Q, t] :: C}{\alpha[f_i \mapsto E[f_k] :: Q, t] :: C \to_\parallel \alpha[f_i \mapsto E[s] :: Q, t] :: C}$$

$$\text{UPDATE-AO}\quad \frac{\gamma \notin \mathrm{dom}(C) \cup \{\alpha\} \qquad \mathit{noFV}(\varsigma(x, y)s) \qquad \beta[R, t'] \in \alpha[f_i \mapsto E[\beta.l := \varsigma(x, y)s] :: Q, t] :: C}{\alpha[f_i \mapsto E[\beta.l := \varsigma(x, y)s] :: Q, t] :: C \to_\parallel \alpha[f_i \mapsto E[\gamma] :: Q, t] :: \gamma[\varnothing, t'.l := \varsigma(x, y)s] :: C}$$

TABLE V. ASP_{FUN} SEMANTICS

C'. Since income is specified as H those two configurations can be considered as α-indistinguishable (with respect to β_1). Attacker α sees no difference between the two. In the fallacious example, however, he'd notice a difference when calling the quicksort algorithm that implicitly drafts information from income through ord: here, C and C' would be distinguishable since β_1.ord is 0 in C and 1 in C'.

E: Noninterference Proof

Theorem 2 (Soundness) For any well-typed configuration, we have noninterference with respect to $\alpha \in C$, i.e.,

$$\left(\begin{array}{c} \vdash C : \langle \Gamma_{\mathit{act}}, \Gamma_{\mathit{fut}}, \mathit{sec}\rangle \\ \alpha[Q, a] \in C \end{array}\right) \Rightarrow \alpha\text{-noninterference } C \ \mathit{sec}\,.$$

Proof:
Let C_1 be another arbitrary but fixed configuration such that $C \sim_\alpha C_1$. This means that for any $\beta \in \mathrm{dom}(C)$, if $\beta \in$ visibility range of α – we have that $\sigma(\beta) \in \mathrm{dom}(C_1)$ and $C(\beta) =_{\sigma,\tau} C_1(\sigma(\beta))$ for some τ and σ. That is, aside differently named futures (and active object references) these two activities are structurally the same and contain the same values. For the sake of clarity of the proof exposition, we leave the naming isomorphism implicit, i.e. use the same names, e.g., β, f_k, for both sides, i.e., for $\beta, \sigma(\beta)$ and $f_k, \tau(f_k)$. Note, that the type of the configurations C' and C_1' is in some cases an extension to the types of C and C_1, as described in Preservation (Theorem 1). The proof is an induction over the reduction relation combined with a case analysis whether an arbitrary $\beta \in \mathrm{dom}(C)$ is in the visibility range of α or not. If, for the first case, $C \to_\parallel C'$ by some reduction according to the semantics rules in the part of the configuration that is *not visible* to α, then for most cases trivially no change becomes visible by the transition to C': for any local reduction, this is the case since the visibility relation is unchanged. Hence, "invisible" objects remain invisible. If $C \sim_\alpha C_1$ and $C \to_\parallel C'$, then also $C' \sim_\alpha C_1$ whereby we trivially have the conclusion since $C_1 \to_\parallel^* C_1$ (in zero steps). This observation is less trivial for the rules REQUEST and ACTIVE where new elements, futures and activities, respectively, are created. In the case of REQUEST, let $\beta[f_k \mapsto E[\gamma.l(t)] :: Q_\beta, t_\beta] \in C$ and $\gamma[Q_\beta, t_\gamma] \in C$ with β in the α-invisible part and γ visible to α. The fact that there are no side effects provides that request f_m created in the request step in γ, i.e. $\gamma[f_m \mapsto t_\gamma.l(t) :: Q_\gamma, t_\gamma]$ in C', is not α-visible. Similarly, if a new activity γ is created from a method in β according to ACTIVE, then γ will not be in the visibility range of α since β was not visible to α by Definition 3.1 of the visibility relation. Thus, for β not visible to α, if $C \sim_\alpha C_1$ and $C \to_\parallel C'$, then also $C' \sim_\alpha C_1$ and the conclusion holds again because $C_1 \to_\parallel^* C_1$. This closes the case of *non-α-visible* reductions. If $C \to_\parallel C'$ by some reduction in the *α-visible* part, we need to consider all cases individually as given by the induction according to the semantics rules. If $C \to_\parallel C'$ by semantics rule REQUEST then C must have contained $\beta[f_k \mapsto E[\gamma.l(t)] :: Q_\beta, t_\beta]$ and $\gamma[Q_\gamma, t_\gamma]$ for some β, f_k, and γ. Hence, $\beta[f_k \mapsto E[f_m] :: Q_\beta, t_\beta]$ and $\gamma[f_m \mapsto t_\gamma.l(t) :: Q_\gamma, t_\gamma]$ in C' for some new future f_m. Since β is α-visible, so is γ by definition of visibility (since f_k was created from t_β, t_β must have an L-method containing γ). By confinement, f_m has global level δ_β and l is L. Since $C \sim_\alpha C_1$, and β, f_k, γ visible to α, we have (up to isomorphism of names) that $\beta[f_k \mapsto E[\gamma.l(t)] :: Q_\beta, t_\beta]$ and $\gamma[Q_\gamma, t_\gamma]$ in C_1. Therefore, we can equally apply the rule REQUEST to C_1 to obtain that C_1' contains $\beta[f_k \mapsto E[f_m] :: Q_\beta, t_\beta]$ and $\gamma[f_m \mapsto t_\gamma.l(t) :: Q_\gamma, t_\gamma]$. In C_1, γ is also α-invisible and l is typed L as well. Now, the α-visible parts of C_1' are equal to the ones of C' apart from the new future f_m. However, based on the future bijection τ that exists due to indistinguishability between C and C_1 we can extend this for f_m to a bijection τ'. In addition, by preservation, C' as well as C_1' are well-typed whereby finally $C' \sim_\alpha C_1'$ and this finishes the REQUEST-case. Another, also less obvious case for new elements in the α-visible part, is the one for ACTIVE. However, here we have a very similar situation as in the REQUEST case. If, in C, there is some $\beta[f_k \mapsto E[\mathit{Active}(t)] :: Q_\beta, t_\beta]$, we also have β alike in C_1, whereby we get in the next step – according to rule ACTIVE – $\beta[f_k \mapsto E[\gamma] :: Q_\beta, t_\beta] \parallel \gamma[\varnothing, t]$ in C' replacing the previous β. We can also apply ACTIVE in C_1 so that $\beta[f_k \mapsto E[\gamma] :: Q_\beta, t_\beta] \parallel \gamma[\varnothing, t]$ in C' and C_1' as well instead of just the old β. Indistinguishability is preserved since a bijection σ' exists as extension to σ to the new activity γ and by preservation again C' and C_1 remain well-typed. We are finished with the case for ACTIVE since $C' \sim_\alpha C_1'$. The other cases, corresponding to the remaining semantics rules, are of very a similar nature. Thus, the second part of α-visible parts of the configurations C and C_1 is also finished and completes the proof of the theorem. □

Processing the Text of the Holy Quran: a Text Mining Study

Mohammad Alhawarat
Department of Computer Science
College of Computer
Engineering and Sciences
Salman Bin Abdulaziz University
Al-Kharj, Kingdom of Saudi Arabia

Mohamed Hegazi
Department of Computer Science
College of Computer
Engineering and Sciences
Salman Bin Abdulaziz University
Al-Kharj, Kingdom of Saudi Arabia

Anwer Hilal
Department of Computer Science
Preparatory Year Deanship
Salman Bin Abdulaziz University
Al-Kharj, Kingdom of Saudi Arabia

***Abstract*—The Holy Quran is the reference book for more than 1.6 billion of Muslims all around the world Extracting information and knowledge from the Holy Quran is of high benefit for both specialized people in Islamic studies as well as non-specialized people. This paper initiates a series of research studies that aim to serve the Holy Quran and provide helpful and accurate information and knowledge to the all human beings. Also, the planned research studies aim to lay out a framework that will be used by researchers in the field of Arabic natural language processing by providing a "Golden Dataset" along with useful techniques and information that will advance this field further. The aim of this paper is to find an approach for analyzing Arabic text and then providing statistical information which might be helpful for the people in this research area. In this paper the holly Quran text is preprocessed and then different text mining operations are applied to it to reveal simple facts about the terms of the holy Quran. The results show a variety of characteristics of the Holy Quran such as its most important words, its wordcloud and chapters with high term frequencies. All these results are based on term frequencies that are calculated using both Term Frequency (TF) and Term Frequency-Inverse Document Frequency (TF-IDF) methods.**

Keywords*—Holy Quran; Text Mining; Arabic Natural Language Processing

I. Introduction

The Holy Quran is the reference book for more than 1.6 billion of Muslims all around the world. Extracting information and knowledge from the Holy Quran is of high benefit for both specialized people in Islamic studies as well as none-specialized people. The Holy Quran is the word of God and hence needs careful handling when processed by automated methods of machine learning, natural language processing and artificial intelligence. The language of the Holy Quran is Arabic which is known to be one of the challenging natural languages in the field of natural language processing and machine learning. This is due to some of its special characteristics such as diacritic, multiple derivations of words, complicated Diglossia and others [1], [2], [3], [4]. These make dealing with Arabic language a challenging task when applying machine learning and artificial intelligence techniques.

Few research studies have considered the Arabic text of Quran [5], [6], [7], [8], instead many studies deal with the translations of the meaning of the words of the holy Quran [9], [10], [11], [12], [13], [14]. Kais and his colleagues have created an open source Quranic corpus [15] using both arabic words as well as translations of these words.

To the best of our knowledge, there is no research study that analyzed the Arabic text of the holy Quran using text mining techniques the way it is done in this paper. The aim of this paper is to find an approach for analyzing Arabic text and then providing statistical information might be helpful for the people in this research area. Also, this study aims at providing a framework for future studies in this field of study. The paper used the holly Quran to achieve these aims; first the holly Quran text is preprocessed and then different packages of *R* has been used such as: `tm and RWeka`.

It is important here to stress that this is not a religious study, instead it is an automated study that gives statistical results. These results in no way are accepted until they are approved by Islamic scholars.

The rest of the paper is organized as the following: Section II is dedicated to explain the process of preparing the text of the Holy Quran, in section III experiments that are applied to the text of the Holy Quran are explained, in section IV the results that are obtained in the paper are discussed, and finally section V concludes the paper.

II. Preparing Text

The holy Quran has around 78 thousand words. These words are grouped into verses. A set of verses are grouped into: parts, chapters, group (Hizb)or Hizb quarter.

The text of the Holy Quran has been first downloaded from Tanzil project website[16] which represents an authentic verified source of the holy Quran text. The downloaded file includes the whole text of the Quran without diacritic. The file has been divided semi-automatically into five different set of documents:

- 114 Chapter (Sura),
- 30 Part (Juza),
- 60 Group (Hizb),
- 240 Hizb Quarter or

- 6236 Verse.

After that the encoding of the files have been converted into `CP1256` because the original encoding of the files are unreadable by *R*.

The files have then been read as a corpus and cleaned by removing stop words. *R* does not support stop word removal for Arabic language, hence a list of around 2000 stop words have been created manually and manipulated from different sources including [17]. Also, *R* does not support stemming for Arabic language, therefore simple cleaning has been applied on the corpus such as normalizing some words by replacing different shapes of the word with its normal form. For example the words:

لله ، والله ، بالله ، فالله ، تالله ، فلله ، اللهم ، آلله

have all been replaced by الله.

Also, the words:

ربنا ، ربهم ، ربكم ، ربك ، ربها ، ربه ، ربي ، برب ، ورب

have all been replaced by رب.

The variations of the previous two words are due to the some of the prefixes and suffixes of the Arabic language. Note that both الله and رب are considered stems rather than roots for the aforementioned variations for both words.

This procedure has been applied to few words because the processing of all shapes of all words (the stemming procedure) is out of the scope of this paper. Although stemming algorithms for Arabic language do exist, but their accuracy still need to be enhanced. For this reason, applying such algorithms is not suitable for the holy Quran as it is the word of God, and hence errors are not tolerated.

After that the corpus have been converted into both Term-Document and Document-Term matrices as both needed for different type of experimentations. The next section illustrate different experimentation applied to both matrices.

III. Experiments

In this section different set of experimentations are carried out on the text of the holy Quran. These experiments are based on the Term matrices that are built according to two selected partitioning methods: Chapters and Parts. These are chosen as examples because using all partitioning methods will produce numerous results and figures.

The experiments will manipulate the text of the holy Quran in order to produce its most frequent terms, wordcloud and clusters.

A. Experiments on Chapters of the holy Quran

In this subsection the text of the holy Quran is studied based on its 114 chapters. Each chapter in the holy Quran talks in general about one theme but it might include different topics. But it is considered the most coherent partitioning methodology.

1) Most Frequent Words: The term-document matrix has been used in one experimentation setup to calculate the frequency of the terms of the holy Quran. There are many frequent terms in the Holy Quran, hence figure 1 depicts the most 30 frequent words. These are calculated using TF measure. Also, the most frequent 30 terms in the holy Quran is calculated based on TF-IDF as shown in figure 2.

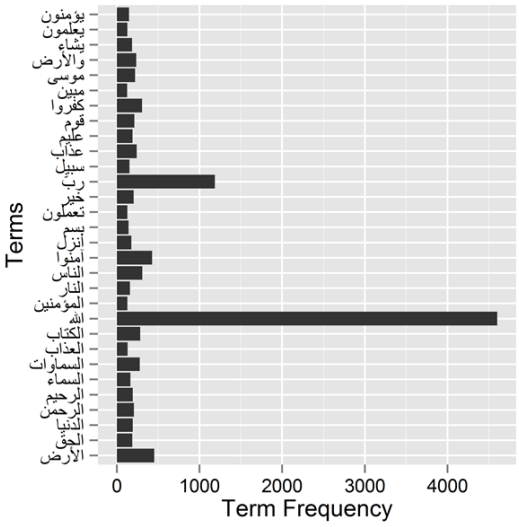


Fig. 1: Most frequent terms in the holy Quran measured by TF (Based on Chapters)

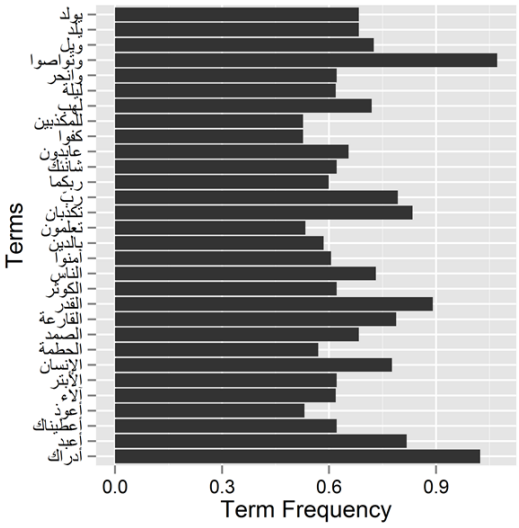


Fig. 2: Most frequent terms in the holy Quran measured by TF-IDF (Based on Chapters)

2) Word Cloud: It is important for specialized as well as non-specialized people in Islamic studies to visualize the words of the Holy Quran. Figures 3- 4 show the wordcloud of the Holy Quran for the most frequent 100 words measured using TF and TF-IDF measures respectively.

Fig. 3: Wordcloud for the most frequent 100 words in the holy Quran measured by TF (Based on Chapters)

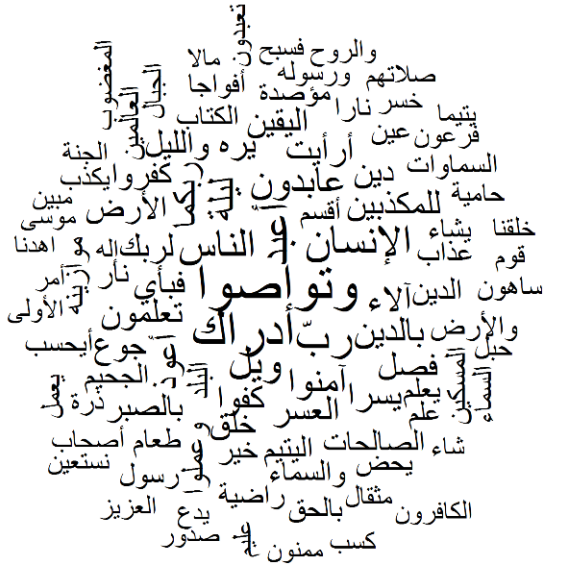


Fig. 4: Wordcloud for the most frequent 100 words in the holy Quran measured by TF-IDF (Based on Chapters)

It's also important to visualize the wordcloud of the holy Quran based on bi-gram, tri-gram and four-gram words. These appear in figures 5- 7.

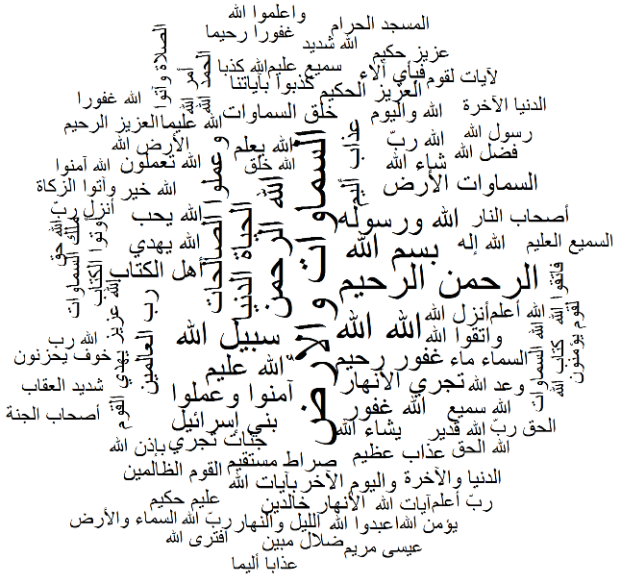


Fig. 5: Wordcloud for the most frequent 100 bi-gram terms (Based on Chapters)

Fig. 6: Wordcloud for the most frequent 100 tri-gram terms (Based on Chapters)

Fig. 7: Wordcloud for the most frequent 100 four-gram terms (Based on Chapters)

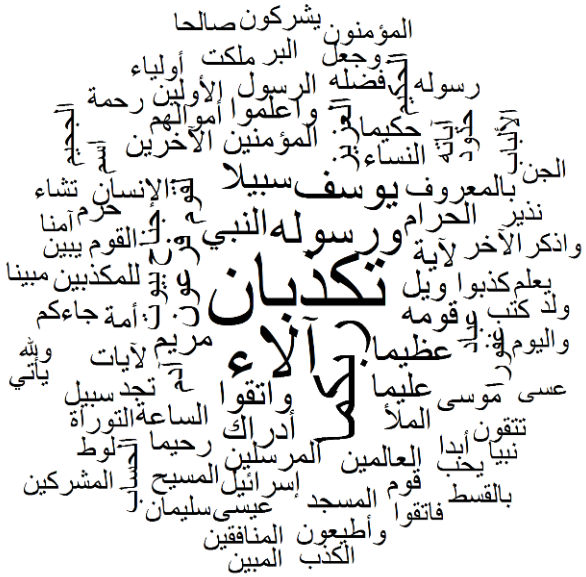


Fig. 9: Wordcloud for the most frequent 100 words in the holy Quran measured by TF-IDF (Based on Parts)

B. Experiments Based on Parts of the holy Quran

The same experimentation setups that are used for Chapters of the holy Quran are applied on the Parts of the holy Quran and samples of the results are depicted in figures 10- 11.

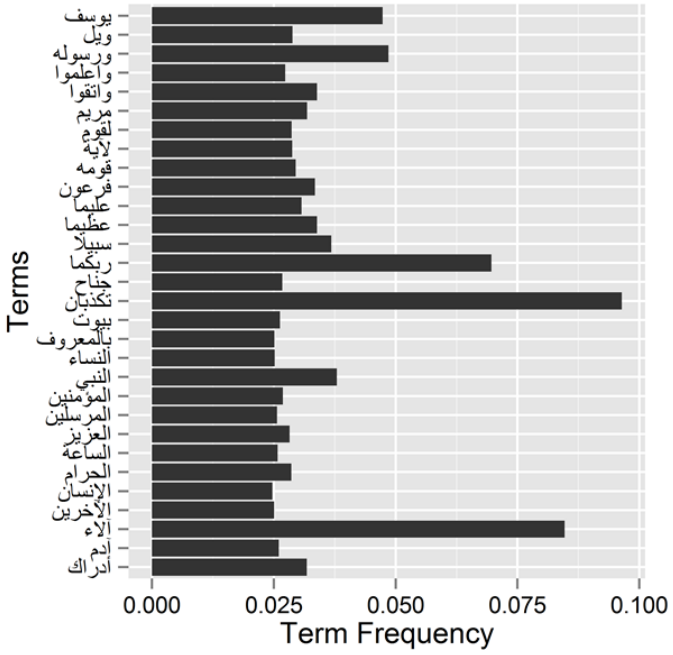


Fig. 8: Most frequent terms in the holy Quran measured by TF-IDF (Based on Parts)

C. Experiments Based on Document-Term Matrix

In this subsection a different experimentation setup is carried out using document-term matrix based on both chapters and parts of the holy Quran.

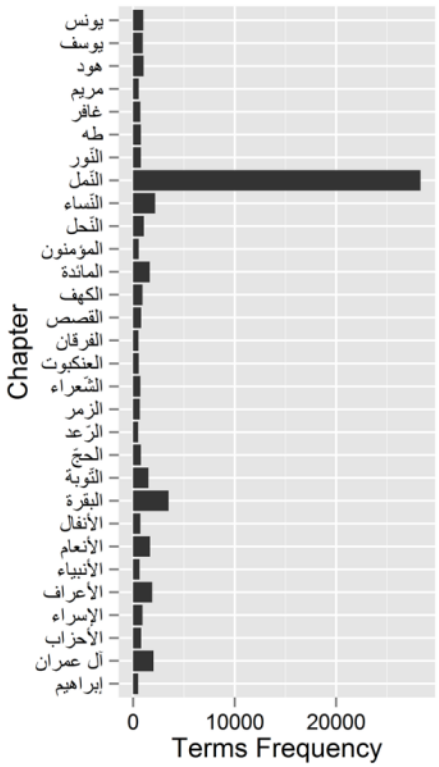


Fig. 10: Term Frequencies based on the Chapters of the holy Quran with Term Frequency larger than 500

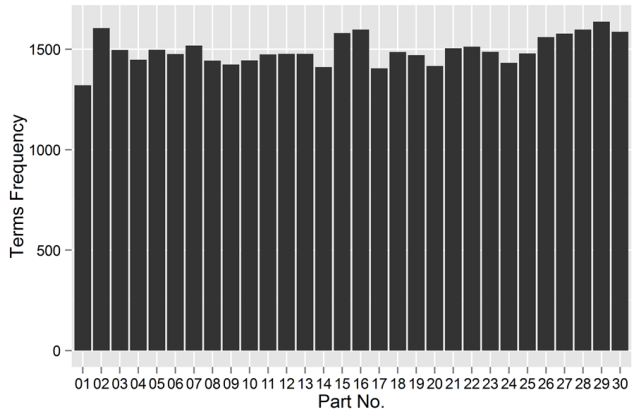


Fig. 11: Term Frequencies based on the Parts of the holy Quran

IV. Discussion

The results that are obtained based on chapters are different from those obtained when parts are used as documents. This is clear when comparing figure 2 with figure 8 where the most frequent terms are different. This is because TF-IDF measure is used; which depends on the ratio of number of all documents to the number of documents a specific term appears on. However, when TF measure is used then there are no differences. That is due to the fact that term-document matrix is used and TF-IDF measure is based on documents compared to TF measure which depends on terms solely.

The bi-grams, tri-grams and four-grams that appear in figures 5- 7 reveal the most frequent n-gram terms in the holy Quran. These terms are of great benefit for future work that is related to semantic search. Notice that these terms are important because stop words are removed before they are extracted.

One important result in this paper is that the partitioning methods affect the distribution of the terms of the holy Quran and there frequencies. This appears when figure 10 and figure 11 are compared. Notice that the term frequencies for Parts are almost equal where the term frequencies for Chapters dramatically vary from one chapter to another. For example, chapter number 19(”The Ant” chapter) contains term frequency of around 28000, where the term frequency of the terms for rest of the chapters of the holy Quran doesn’t reach 5000.

There are many applications for the most frequent terms obtained in this paper. For example, frequent terms presented in figure 1 and figure 2 are calculated using TF and TF-IDF measures respectively. These calculations were based on chapters of the holy Quran; however results that are shown in figure 8 were calculated based on Parts of the holy Quran. Choosing the right application depends on both the document-partitioning method as well as the calculation measure. The first might affect the precision and quality of results of a specific application; and hence need to be tested in a specific application. However, the calculation measure might affect choosing the right application, for example if semantic search or clustering applications are chosen then it is more suitable to use terms produced using TF measure. This is due to the fact that it will give more weight for terms that are repeated most in all documents. On the other hand, if topic modelling application is chosen, then it is more appropriate to use terms calculated based on TF-IDF measure. This is because using it will return terms that are more important in a specific document and less important in other documents.

The results obtained in this paper could be improved if a stemming algorithm is used. This is because using stems instead of the original words will give more accurate results as known in information retrieval field of study; where using stems will increase the precession. For Arabic language, stemming algorithms are still immature and have high error rates. For the holy Quran, error rates was calculated and it is in the range of 22-55%. For more details please see [18], [19], [20]. Such error rates are not acceptable when analyzing the text of the holy Quran because it is the word of God and errors will change the meanings of its words.

V. Conclusion

This study aims to layout a framework for future work that is related to the application of natural language processing, data mining and text mining to the text of the holy Quran. This is done by initially preprocessing the text of the holy Quran and by considering the different possible partitioning. Choosing one document portioning of the holy Quran affects the resulted Frequent Terms of the holy Quran. Moreover, if TF-IDF weighting is used, then the resulted Frequent Terms are also change. Graphical representation of the main terms of the holy Quran is depicted using term-frequencies plot and wordcloud plots including bi-grams, tri-grams and four-grams.

One important result of this paper is that chapter ”The Ant” contains the largest term frequency in the holy Quran; this could be of benefit for future work when analyzing and clustering the text of the holy Quran. Also such a result might be of interest for scholar and specialized people in Islamic studies.

Another important result in this study is that a list of frequent terms are suggested to be used in different applications: on one hand for, those terms calculated based on TF measure they might give good results for semantic search and clustering; on the other hand, terms that are calculated based on TF-IDF measure are more suitable for topic modelling.

This study constitutes the first phase in a large project that aims at advancing the research in the field of arabic natural language and more specifically in exploring the text of the holy Quran. Therefore, the results illustrated in this paper represent a simple sample of what could be obtained from the analysis of the text of the holy Quran.

Although the results of this study are interesting, however it is based on the original words of the holy Quran. More accurate results will be obtained if an efficient stemming algorithm is used. Unfortunately, there is no accurate known stemming algorithm exists for Arabic language due to the challenges that are faced in processing Arabic language.

Future work may include preprocessing the text of the holy Quran with efficient and accurate algorithm that might give words like stems as light stemmers algorithms do. If such algorithm is developed then further study on the text

of the holy Quran will be carried out to extract knowledge and important information that is useful to all humanity using machine learning techniques.

ACKNOWLEDGMENT

This project was supported by the deanship of scientific research at Salman bin Abdulaziz University under the research project number 109/T/33.

REFERENCES

[1] M. DIAB and N. HABASH, "Arabic dialect tutorial," in *In Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics (NAACL07)*, 2007, pp. 29–34.

[2] A. Farghaly and K. Shaalan, "Arabic natural language processing: Challenges and solutions," vol. 8, no. 4, pp. 14:1–14:22, Dec. 2009. [Online]. Available: http://doi.acm.org/10.1145/1644879.1644881

[3] N. Y. Habash, *Introduction to Arabic Natural Language Processing*, G. Hirst, Ed. Morgan and Claypool Publishers, 2010.

[4] M. Saad and W. Ashour, "Arabic morphological tools for text mining," in *6th International Symposium on Electrical and Electronics Engineering and Computer Science, European University of Lefke, Cyprus, 2010*, 2010, p. 112117.

[5] I. Ali, "Application of a mining algorithm to finding frequent patterns in a text corpus: A case study of the arabic," *International Journal of Software Engineering and Its Applications*, vol. 6, pp. 127–134, 2012.

[6] M. Al-Yahya, H. Al-Khalifa, A. Bahanshal, I. Al-Odah, and N. Al-Helwah, "An antological mdoel for representing semantic lexicons: An application on times nouns in the holy quran," *The Arbaian Journal for Science and Engineering*, vol. 35(2c), pp. 21–37, 2010.

[7] K. Dukes, E. Atwell, and N. Habash, "Supervised collaboration for syntactic annotation of quranic arabic," *Language Resources and Evaluation*, vol. 47, no. 1, pp. 33–62, 2013. [Online]. Available: http://dx.doi.org/10.1007/s10579-011-9167-7

[8] M. H. Panju, "Statistical extraction and visualization of topics in the qur'an corpus," Master's thesis, University of Waterloo, 2014.

[9] N. Ismail, N. Rahman, Z. Bakar, and T. Sembok, "Terms visualization for malay translated quran documents," in *International Conference on Electrical Engineering and Informatics, Bandung, Indonesia*, 2007, pp. 554–557.

[10] M. S. Hikmat Ullah Khan, Syed Muhammad Saqlain and M. Sher, "Ontology-based semantic search in holy quran," vol. 2, no. 6, pp. 562–566, 2013.

[11] N. Shahzadi, A. ur rahman, and M. J. Sawar, "Semantic network based classifier of holy quran," *International Journal of Computer Applications*, vol. 39, no. 5, pp. 43–47, February 2012.

[12] M. Shoaib, M. Nadeem Yasin, U. Hikmat, M. Saeed, and M. Khiyal, "Relational wordnet model for semantic search in holy quran," in *International Conference on Emerging Technologies, 2009. ICET 2009.*, Oct 2009, pp. 29–34.

[13] A. A. Aliyu Rufai Yauri, Rabiah Abdul Kadir and M. A. A. Murad, "Quranic verse extraction base on concepts using owl-dl ontology." vol. 6, no. 23, pp. 4492–4498, 2013.

[14] M. Yunus, R. Zainuddin, and N. Abdullah, "Semantic query for quran documents results," in *Open Systems (ICOS), 2010 IEEE Conference on*, Dec 2010, pp. 1–5.

[15] K. Dukes. (2014, Dec.) Quranic arabic corpus. [Online]. Available: http://corpus.quran.com/

[16] Tanzil.net. (2014, Oct.) Tanzil quran text download @ONLINE. [Online]. Available: http://tanzil.net/download/

[17] T. Zerrouki. (2014, Nov.) Arabic stop words @ONLINE. [Online]. Available: http://sourceforge.net/projects/arabicstopwords/

[18] M. Sawalha and E. Atwell, "Comparative evaluation of arabic language morphological analysers and stemmers," in *Coling 2008: Companion volume: Posters*. Coling 2008 Organizing Committee, 2008, pp. 107–110.

[19] N. Thabet, "Stemming the qurán," in *Workshop on Computational Approaches to Arabic Script-based Languages*. Coling 2008 Organizing Committee, 2008, pp. 28–31.

[20] R. J. R. Yusof, R. Zainuddin, M. S. Baba, and Z. M. Yusoff, "Quránic words stemming," *Arabian Journal for Science and Engineering*, vol. 35, p. 38, 2010.

SOCIA: Linked Open Data of Context behind Local Concerns for Supporting Public Participation

Shun Shiramatsu*, Tadachika Ozono* and Toramatsu Shintani*
*Department of Computer Science and Engineering, Graduate School of Engineering, Nagoya Institute of Technology,
Gokiso-cho, Showa-ku, Nagoya 466-8555, Japan

***Abstract*—To address public concerns that threat the sustainability of local societies, supporting public participation by sharing the background context behind these concerns is essentially important. We designed a *SOCIA ontology*, which was a linked data model, for sharing context behind local concerns with two approaches: (1) structuring Web news articles and microblogs about local concerns on the basis of geographical regions and events that were referred to by content, and (2) structuring public issues and their solutions as public goals. We moreover built a *SOCIA dataset*, which was a linked open dataset, on the basis of the SOCIA ontology. Web news articles and microblogs related to local concerns were semi-automatically gathered and structured. Public issues and goals were manually extracted from Web content related to revitalization from the Great East Japan Earthquake. Towards more accurate extraction of public concerns, we investigated feature expressions for extracting public concerns from microblogs written in Japanese. To address a technical issue about sample selection bias in our microblog corpus, we formulated a metric in mining feature expressions, i.e., bias-penalized information gain (BPIG). Furthermore, we developed a prototype of a public debate support system that utilized the SOCIA dataset and formulated the similarity between public goals for a goal matching service to facilitate collaboration.**

***Keywords*—Semantic Web; social computing; natural language processing; linked open data; e-Participation**

I. Introduction

Japanese regional societies currently face complicated and ongoing social issues or concerns, e.g., dwindling birth rates, an aging population, public finance problems, disaster risks, dilapidated infrastructures, and radiation pollution that threaten the sustainability of societies. The coverage of government services is expected to decrease along with an escalation in these concerns. Some Japanese researchers regard such troubling situations as “a front-runner of emerging issues”[1]. To address these concerns, supporting public participation by sharing background context behind these concerns is essentially important.

We have aimed to develop a Web platform to support public participation, which provides a function for sharing background context behind local concerns [2], [3], [4]. Since citizens who have beneficial awareness or knowledge are not always experts on relevant social concerns, background context needs to be shared to reduce barriers to public participation. It is difficult to participate in addressing concerns without background context. Linked open data (LOD)[5], which are semantically connected data on the basis of universal resource identifiers (URIs) and the resource description framework (RDF), play an important role in fostering open government [6]. To increase transparency and participation in regional communities, it is important for citizens, government officials, and experts to share public concerns. Background context should be structured and open to facilitate the assessment and sharing of public concerns. The LOD framework is suitable for structuring such background contexts and concerns. The structure of public concerns is an important context when building consensus. We have called the process of structuring public concerns “concern assessment”.

We designed a linked data model and built an LOD dataset, which were called *Social Opinions and Concerns for Ideal Argumentation (SOCIA)*, to share the context behind local concerns. The data model of *SOCIA ontology* was designed with two approaches. The first was attained by structuring Web news articles and microblogs about local concerns on the basis of geographical regions and events that were referred to by the content. The second was attained by structuring public issues and their solutions as public goals. We moreover built a *SOCIA dataset*, which was a linked open dataset (LOD), on the basis of the SOCIA ontology. Japanese local news articles, microblog posts, and minutes of city council meetings are semi-automatically structured on the basis of geographical regions and events. The SOCIA dataset also included public issues and goals that were manually extracted from news articles.

Furthermore, we preliminarily investigated feature expressions to extract public concerns from microblogs written in Japanese. The feature expressions were mined from a corpus consisting of microblogs about public concerns (positive examples) and microblogs about irrelevant to public concerns (negative examples). We addressed a technical issue about the sample selection bias in the positive examples, i.e., there were unsuitable feature expressions that were frequently used by only one specific person.

The rest of the paper is organized as follows. Section II presents conventional works related to e-Participation. The SOCIA ontology is described in Section III. Section IV describes the SOCIA dataset built by semi-automatically structuring Web content related to local concerns and manually structuring public issues and goals extracted from Web content. Section V explains how Japanese feature expressions for extracting public concerns from microblogs were mined with a corpus-based approach. Section VI describes applications of the SOCIA dataset and Section VII concludes the paper.

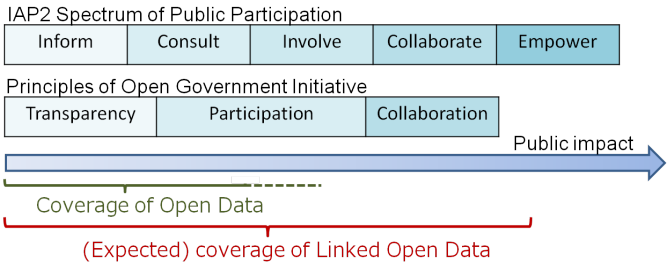


Fig. 1: Expected coverage of Linked Open Data on the spectrum of public participation

II. Related Works

A. Public Participation and Open Data

The International Association for Public Participation (IAP2) and the Obama administration's Open Government Initiative (OGI) have presented similar stages for public participation, i.e., the Spectrum of Public Participation[7] and the Principles of Open Government[8] shown in Figure 1. The gradation in the figure represents the public impact of each stage. The figure also indicates the expected coverage of the use of LOD. Open data generally contributes to transparency, i.e., to the first stage. However, non-linked open data (e.g., CSV table data) generally lack interoperability. LOD is expected to be able to also contribute to the higher/collaborative stages because semantic links compliant with RDF increase the interoperability of data and help us to reuse data for inter-organizational collaboration. Contextual information provided by the semantic links provides the potential for developing social Web services to facilitate public collaboration.

Over 40 countries currently provide open data portals.[1] The number of open data portals has been increasing since 2009. An open data portal by the Japanese government, data.go.jp, was also launched in 2014. One hundred local governments (14 prefectures and 86 municipalities) in Japan also provide their open government data as of Feb. 2015[2].

B. Modeling Public Debate and Participation

Providing background information related to public debate is important in order to support concern assessment. In view of this, argument visualization is an effective approach for supporting eParticipation [9]. Jeong et al. visualized the difference in cognition for several topics among participants in public debates using the co-occurrence of terms [10]. Visualizing an overview of public debate is also effective for grasping the background. Several argument visualization tools currently exist [11]: Compendium [12], Cohere [13], MIT Deliberatorium [14], Araucaria [15], Discourse Semantic Authoring [16], [17], etc. Typically, these tools produce "box and arrow" diagrams in which premises and conclusions are formulated as statements [18].

Within the context of LOD and the semantic Web, the Talk of Europe project proposed a linked data model to structure public debate [19]. Their data model focuses on transcripts of the plenary meetings of the Talk of Europe. Within a broader context, Porwol et al. designed an e-Participation ontology, which was a semantic model of e-Participation [20]. The ontology contained classes of `epart:Project`, `epart:Platform`, and `epart:DemocraticProcess`.

[1] http://www.data.gov/opendatasites
[2] http://fukuno.jig.jp/2013/opendatamap (in Japanese)

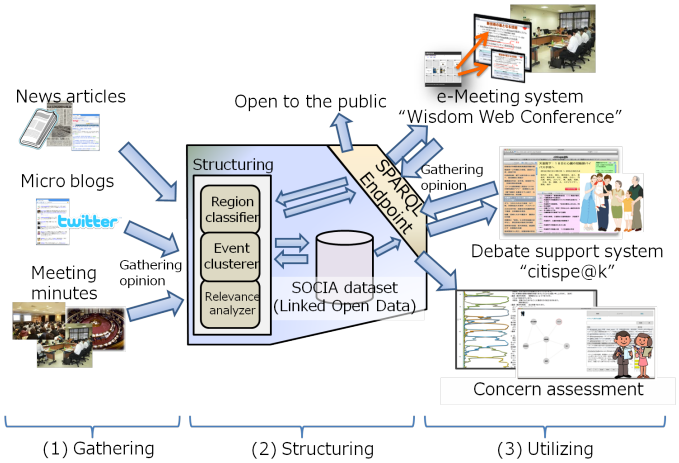


Fig. 2: Outline of O_2, e-Participation Web Platform

III. Designing SOCIA Ontology

This section describes the design of the SOCIA ontology to structure Web news articles and microblogs about local concerns on the basis of geographical regions and events that are referred to by content, and to structure public issues and their solutions as public goals.

A. Structuring Web Content about Local Concerns

To design a data model for sharing background context behind local concerns, we consider applications of the dataset. O_2, an abbreviation for Open Opinion, is our Web platform for citizen participation in debates about regional issues. As shown in Fig. 2, the O_2 platform has three stages. In stage (1), the mining and pre-processing system crawls the Web and gathers information from news articles, microblogs, and meeting minutes that can be used for debates. In stage (2),

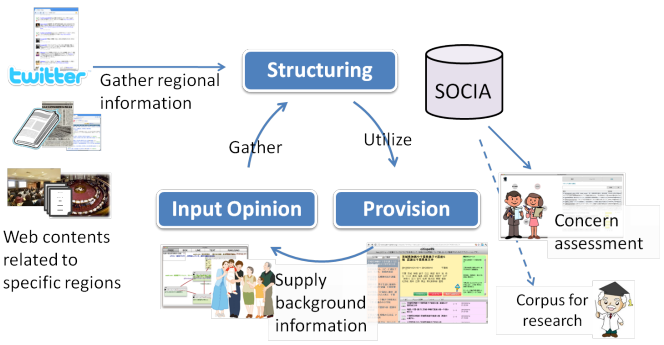


Fig. 3: Cycle of utilizing regional information for e-Participation

the system geographically classifies the gathered contents and clusters them by event. Relevant information is then structured and stored in the SOCIA dataset in accordance with the SOCIA ontology as openly published Linked Open Data. In stage (3), the structured information is used for public participation, i.e., debate support, concern assessment, etc.

The cycle of utilizing regional information in SOCIA for eParticipation is illustrated in Fig. 3. To help citizens understand public concerns and express their opinions, background information needs to be provided because most citizens are not experts about diversified public concerns. The opinions expressed can also be utilized as background information after being structured in the SOCIA dataset. For Web contents (e.g. news articles, blogs, and tweets) to be used as background information, they need to be classified by region and then presented to citizens in an understandable way. Our platform and ontology can be used to structure the URLs of Web contents and then link them with regional issues.

The SOCIA dataset is openly published on the Web using the SOCIA ontology,[3] designed using Web Ontology Language (OWL) as shown in Fig. 4. Through this process, eParticipative

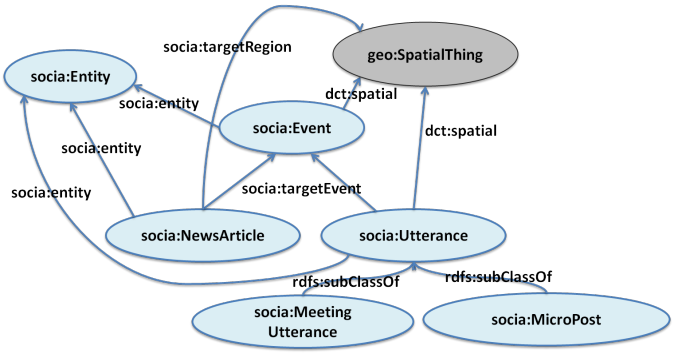


Fig. 4: Core classes for structuring regional information in SOCIA ontology

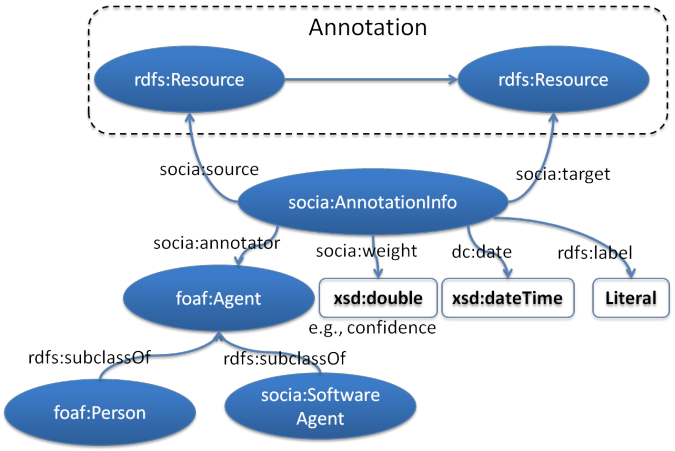


Fig. 5: AnnotationInfo: meta-context information related to property annotation

[3]http://data.open-opinion.org/socia-ns

data becomes re-usable and transparent.

Text mined from the Web is structured in the form of events by region, which are then used as discussion seeds to further build the SOCIA dataset. Citizens then create discussion topics out of each seed, e.g., a cluster of news articles related to the same event, and input their opinions by using the system, among other functionalities.

To improve the structuring accuracy, the history of how the LOD properties were annotated (e.g., which algorithm, which parameter, by whom is needed) because the automatic structuring by Sophia has an inherent error of a few percent. To maintain the annotation history, we defined the AnnotationInfo class, as shown in Fig. 5. Such meta-context information is necessary when the data set is used as a corpus for research on natural language processing.

B. Structuring Public Issues and Goals

Public collaboration and consensus building between stakeholders are essential to enable revitalization from disasters, e.g., the Great East Japan Earthquake. Collaboration between multiple agents generally requires the following conditions:

- Similarity of the agents' goals or objectives
- Complementarity of the agents' skills, abilities, or resources

As the first step, this study focuses on the similarity of the goals. Sharing a data set of public goals can help citizens, who have similar goals, build consensus and collaborate with one another.

We focus on the following three problems related to public collaboration.

1) Citizens cannot easily find somebody whose goals are similar to their ones.
2) Stakeholders who have similar goals occasionally conflict with one another when building consensus because subgoals are sometimes difficult to be agreed on even if the final goal is generally agreed on.
3) A too abstract and general goal is hard to be contributed collaboratively.

We presume that the hierarchies of goals and subgoals play important roles to address these problems. First, the hierarchical structure can make methods of calculating the similarity between public goals more sophisticated. The hierarchy provides rich context to improve retrieval of similar goals. If the data set of public goals had only short textual descriptions without hierarchical structures, calculating the similarity between goals would be difficult and the recall ratio in retrieving similar goals would be lower. Second, visualizing the hierarchies is expected to support people in conflict to attain compromises. Third, dividing goals into fine-grained subgoals reduces barriers to participation and collaboration because small contributions to fine-grained subgoals are more easily provided.

Fig. 6 shows an extention of the SOCIA ontology to represent public issues and goals. The classes `socia:Issue` and `socia:Goal` are connected with the `socia:solution` property. These classes are linked with `foaf:Agent` corresponding to participants or stakeholders and with

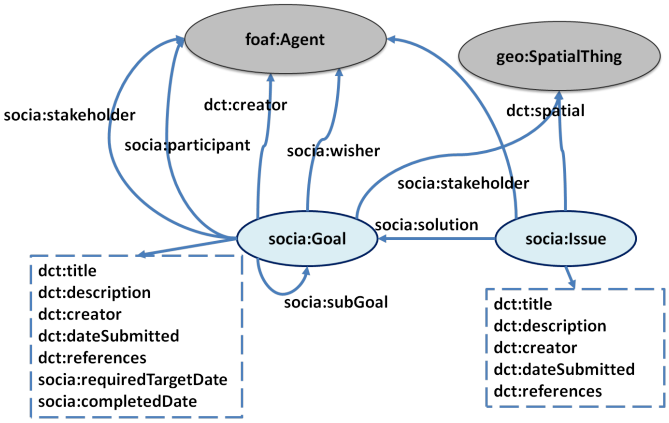


Fig. 6: Core classes for structuring public issues and goals in SOCIA ontology

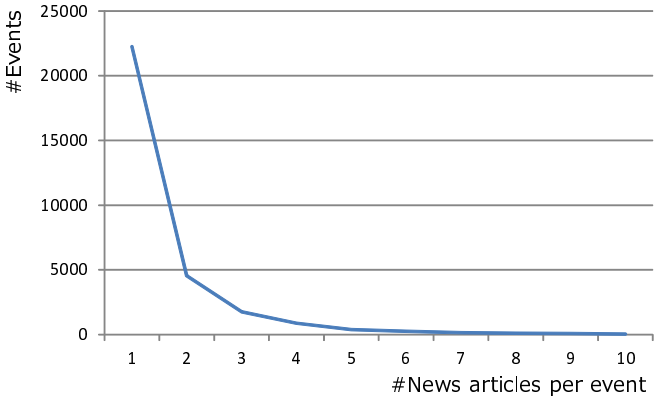


Fig. 7: Distribution of news article counts per event

`geo:SpatialThing` corresponding to geographical regions.

IV. Building SOCIA Dataset

This section describes semi-automatic structuring of Web content on local concerns and manual structuring of public issues and goals.

A. Gathering Web Content about Local Concerns

The system first collects news articles, microblog posts (in this work, tweets), and minutes of city council meeting from the Web along with necessary metadata (dates, emission sources, etc). It then classifies this crawled Web contents by region and filters out contents unrelated to the interests of regional communities or to current events. Next, the system extracts target events from the news articles and microblogs, and links them using the ontology.

Citizens can then add further links to events, news articles, and microblogs, by creating relevant topics and can debate them by inputting their opinions, polling, or sharing further resources. Those resources and new links are also incorporated in the data set, as are the opinions and the discussion. This creates a virtuous cycle in which the intelligent platform, by creating understandable and relevant discussion seeds, involves citizens in eParticipation. The citizens add further data to the data set, making it grow over time, and this data can be used as input again (e.g. for training better learning models and developing better ontologies).

1) Classification by Geographic Region: After the mining, the gathered news articles and tweets are classified geographically (by the 47 prefectures of Japan). To this end, we use Transformed Weight-normalized Complementary Naive Bayes (TWCNB) algorithm [21]. In the classification, the feature vectors for each document consist of the TF*IDF value of morpheme bi-grams. To decide whether contents should be filtered out or not, we use a confidence threshold where the confidence value is defined as the difference between log scores of the highest-ranked class and that of second-ranked class.

We conducted a classification experiment through varying threshold of confidence value, using 8,811 news articles related to Japanese prefectures crawled from Yahoo! Japan News[4] during Jun. 13 to Jul. 12, 2011, and 1,133 ones that do not related to any prefectures. The experimental result showed that the precision is 98.2% and the recall is 98.0% for the optimal threshold [22], [23].

2) Clustering by Events: The SOCIA dataset stored 54,854 news articles, with about 13,000 ones classified as related to a prefectures.[5] The events are extracted as clusters of similar news articles [23]. The similarity between news articles are calculated as a cosine similarity which is weighted by a window function determined by for considering dates/times the news articles were published. As shown in Fig. 7, about 35,000 events were extracted through the clustering of these articles.

B. Manual Extraction of Public Goals from Web News Articles

We built an LOD set[6] by manually extracting public goals from news articles and related documents. The 657 public

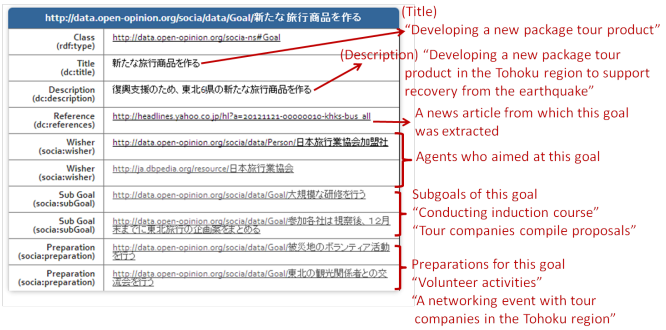


Fig. 8: Instance of public goal: "Developing new package tour product"

[4] http://headlines.yahoo.co.jp/hl?c=loc

[5] The number of news articles stored in SOCIA was counted on Mar. 16, 2012. It has been constantly increasing.

[6] http://data.open-opinion.org/socia/data/Goal?rdf:type=socia:Goal&limit=700 (in Japanese)

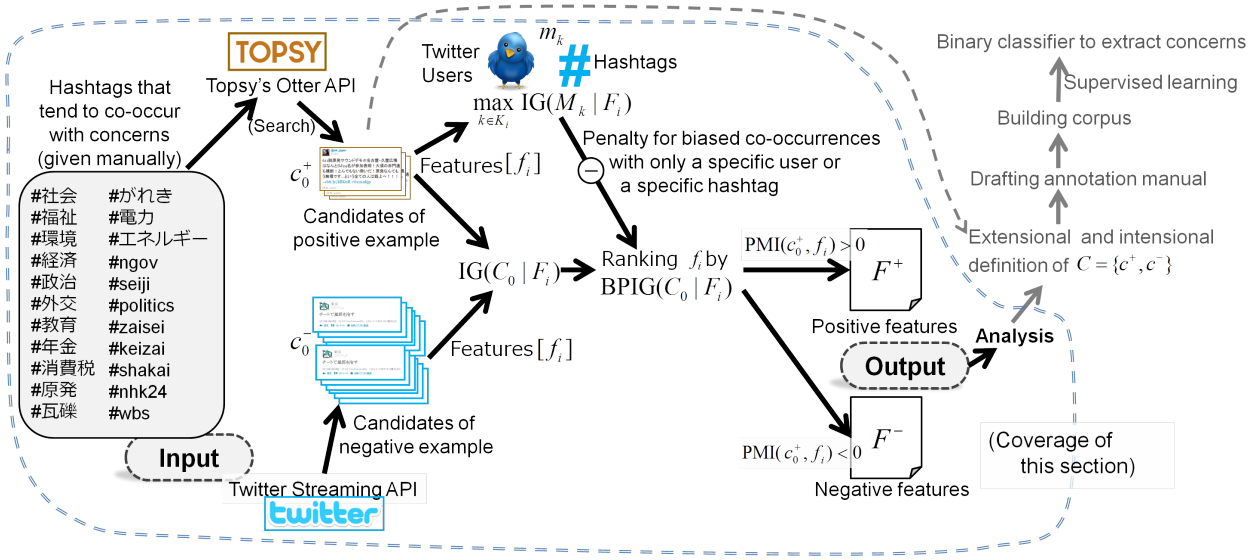


Fig. 9: Processing flow for mining features to extract public concerns

goals and 4349 RDF triples were manually extracted from 96 news articles and two related documents by one human annotator. The most abstract goal that is the root node of the goal-subgoal hierarchy is "revitalization from the earthquake".[7] The subgoals are linked from this goal with the `socia:subgoal` property.

The manually built LOD set can be used for developing a method of calculating the similarities between public goals. It can also be used as example seed data when citizen users input their own goals for revitalization. Fig. 8 shows an instance of a public goal to revitalize the Tohoku region from the Great East Japan Earthquake. This goal of "developing a new package tour product", has a title in Japanese, a description in Japanese, and two subgoal data resources.

This dataset about public goals for revitalization won the 2nd Prize of Dataset Track of the Linked Open Data Challenge Japan 2013[8].

V. Mining Feature Expression to Extract Concerns

Automatic structuring needs to become more accurate with a filter for noisy text to support concern assessment because consumer-generated Web content (e.g., microblogs) frequently contains noise information on the target regions. We aimed to construct a binary classifier between tweets including public concerns and others. To define the boundary between the positive class c^+ (corresponding to public concerns) and the negative class c^- (corresponding to tweets other than public concerns), we investigate approximative examples collected through hashtag search. Figure 9 represents the processing flow for investigating the approximate examples. Firstly, we manually prepare the list of hashtags that may frequently co-occur with public concerns in Japanese tweets: #政治 (politics), #社会 (society), #環境 (environment), and so on. The tweets collected through searching by these hashtags from Topsy's Otter API[9] are regarded as candidates of positive examples. These examples are labeled as class c_0^+, an approximative positive class. However, note that the c_0^+ examples also include noise tweets that are not suitable for concern assessment. Secondly, we gather general tweets from Twitter Streaming API[10]. The ratio of public concern in this set is much less than that in the c_0^+ set. Therefore, these general tweets are regarded as candidates of negative examples and labeled as class c_0^-, an approximative negative class. In this section, we empirically analyze features for classifying tweets into $C_0 = \{c_0^+, c_0^-\}$ towards building a corpus annotated with $C = \{c^+, c^-\}$, a more sophisticated concern definition.

Here, we denote a feature vector of a tweet by $[f_i]_i$. Let $F_i = \{f_i^+, f_i^-\}$ where f_i^+ denotes a label representing that the feature f_i appears in a tweet, and f_i^- denotes a label representing that f_i does not. A feature f_i's significance for extracting c_0^+ tweets can be estimated by the information gain:

$$\mathrm{IG}(C_0|F_i) = \mathrm{H}(C_0) - \mathrm{H}(C_0|F_i), \tag{1}$$

with

$$\mathrm{H}(C_0) = -p(c_0^+)\log p(c_0^+) - p(c_0^-)\log p(c_0^-), \tag{2}$$

$$\mathrm{H}(C_0|F_i) = -p(c_0^+|f_i^+)\log p(c_0^+|f_i^+) - p(c_0^-|f_i^+)\log p(c_0^-|f_i^+) -p(c_0^+|f_i^-)\log p(c_0^+|f_i^-) - p(c_0^-|f_i^-)\log p(c_0^-|f_i^-). \tag{3}$$

The features f_i extracted from c_0^+ tweets with the information gain, however, are biased due to sample selection

[7] `http://data.open-opinion.org/socia/data/Goal/%E9%9C%87%E7%81%BD%E5%BE%A9%E8%88%88` (in Japanese)

[8] http://lod.sfc.keio.ac.jp/blog/?p=2074 (in Japanese)

[9] http://otter.topsy.com/

[10] https://dev.twitter.com/docs/streaming-api

bias dependent on the input hashtags. To address the sample selection bias, we formulate bias-penalized information gain (BPIG) with considering a penalty for biased occurrence of feature f_i as follows:

$$\mathrm{BPIG}(C_0|F_i) = \mathrm{IG}(C_0|F_i) - \alpha \max_{k \in K_i} \mathrm{IG}(M_k|F_i) \quad (4)$$

with

$$K_i = \{k \mid \mathrm{PMI}(m_k, f_i|c_0^+) > 0\} \quad (5)$$

$$\mathrm{PMI}(m_k, f_i|c_0^+) = \log \frac{p(m_k, f_i|c_0^+)}{p(m_k|c_0^+)p(f_i|c_0^+)} \quad (6)$$

$$M_k = \{m_k^+, m_k^-\}, \quad (7)$$

where let m_k^+ be a label representing that m_k, a hashtag or a user, appears in a tweet or is the author of the tweet, m_k^- be a label representing that m_k does not, and $\alpha \in [0, 1]$ be a weight of the penalty term. Here, $\max_{k \in K_i} \mathrm{IG}(M_k|F_i)$ can be regarded as a penalty for f_i that co-occurs only with a particular hashtag or user m_k.

Table I shows the hashtags for gathering c_0^+ tweets from Topsy's otter API. We specified Japanese as the language of gathered tweets in query URLs for the API. Temporal distribution of the 32,844 tweets collected as c_0^+ is shown in Figure 10. The c_0^+ tweets consist mostly of the tweets in the latest months due to the characteristics of time window of the Topsy search.

TABLE I: Hashtags for gathering c_0^+ tweets

Hashtags	#Tweets
#社会 (society)	1,981
#福祉 (welfare)	1,629
#環境 (environment)	1,380
#経済 (economy)	1,985
#政治 (politics)	3,131
#外交 (diplomacy)	986
#教育 (education)	1,865
#年金 (pension)	940
#消費税 (consumption tax)	1,592
#原発 (nuclear plant)	3,129
#瓦礫 (rubble)	2,367
#がれき (rubble)	1,949

Hashtags	#Tweets
#電力 (electricity)	1,020
#エネルギー (energy)	797
#ngov	1,040
#seiji (politics)	4,796
#politics	1,775
#zaisei (finance)	1,014
#keizai (economy)	2,406
#shakai (society)	1,018
#nhk24	1,844
#wbs	289
Total	38,933
Total without duplication	32,844

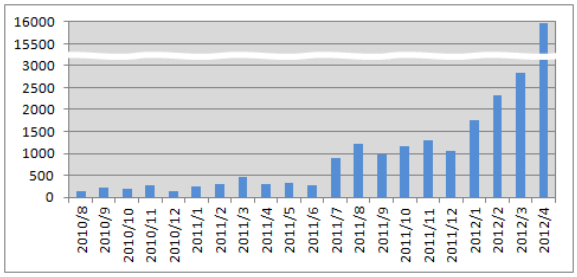


Fig. 10: Temporal distribution of c_0^+ tweets gathered from Topsy Otter's API

TABLE II: Temporal distribution of c_0^- tweets gathered from Twitter streaming API

Duration (JST)	#Tweets
2011-10-16 21:44:25〜23:55:31	49,998
2012-02-20 11:19:25〜15:25:04	49,994
2012-04-14 00:59:15〜07:57:55	49,992
Total	149,984

The c_0^- tweets are gathered from Twitter Streaming API. The ratio of public concerns in c_0^- is predicted to be much less than that in c_0^+. Temporal distribution of the 149,984 tweets collected as c_0^- is shown in Table II. Since we presume that the ratio of c^- is greater than that of c^+, the ratio of c_0^- is also set as greater than that of c_0^+. We conducted an experiment for feature extraction using these 182,828 tweets consisting of c_0^+ and c_0^-. Features representing c_0^+ and c_0^- are extracted with the following procedure:

1) Rank features f_i by $\mathrm{IG}(C_0|F_i)$ and $\mathrm{BPIG}(C_0|F_i)$, respectively.
2) As features for c_0^+, extract high-ranked features f_i, such that $\mathrm{PMI}(c_0^+, f_i) = \log \frac{p(c_0^+, f_i)}{p(c_0^+)p(f_i)} > 0$.
3) As features for c_0^-, extract high-ranked features f_i, such that $\mathrm{PMI}(c_0^+, f_i) < 0$.

In this experiment, we regard morpheme N-grams as features of each tweet. Table III and IV represent the results of feature extraction where let $N = 3$, i.e., in case of morpheme tri-grams. There are some pre-processings before extracting morpheme N-grams; URL strings and user names (starting with @) in tweets are replaced by "[URL]" and "[USER]". Hashtags in tweets are omitted. "[B]" and "[E]" are inserted into beginning and end of a tweet, respectively.

The features for the positive example, c_0^+, are shown in Table III. The features extracted by information gain, which are ranked in the left side of the table, are greatly biased due to the input hashtags. For example, both of "NEWS WEB 24" (a name of TV news program) and "番組で紹介" (introducing it in our program) are dependent on the hashtag #nhk24. In contrast, the features extracted by BPIG in the right sides of the tables are not specific to a particular hashtag of a user. These N-gram features are commonly used for describing public concerns, e.g., expressions for stating fact or question. Table V represents features f_i which have higher penalties for bias, that is, higher $\max_{k \in K_i} \mathrm{IG}(M_k|F_i)$. The result shows that BPIG can appropriately filter out features that co-occurs only with a particular hashtag or user.

The features for the negative example, c_0^-, are shown in Table IV. Both the c_0^+'s features and the c_0^-'s features are needed for classifying the positive examples and the negative ones. The c_0^-'s features can be used for filtering the negative examples as noise tweets. Although in both cases of information gain and BPIG, expressions for greeting or communication are higher ranked, features with higher $p(c_0^+|f_i)$, such as " !! [E] " and " !!! ", are lower-ranked in BPIG than in information gain.

Morpheme N-grams ($N = 2, 3, 4, 5$) extracted as features for c_0^+ can be classified by modality types as shown in Table

TABLE III: Morpheme tri-grams extracted as features representing c_0^+

Ranking by IG			Ranking by BPIG		
Tri-gram f_i	$\mathrm{IG}(C_0\|F_i)$	$p(c_0^+\|f_i)$	Tri-gram f_i	$\mathrm{BPIG}(C_0\|F_i)$	$p(c_0^+\|f_i)$
）[URL][E]	8.38×10^{-3}	0.804	」[URL][E]	4.14×10^{-3}	0.758
』[URL][E]	8.25×10^{-3}	0.843	ている。	1.21×10^{-3}	0.602
:[URL][E]	6.79×10^{-3}	0.845	ているの	9.23×10^{-4}	0.560
...[URL][E]	5.96×10^{-3}	0.751	ています	9.14×10^{-4}	0.539
」[URL][E]	5.51×10^{-3}	0.758	している	9.13×10^{-4}	0.602
。[URL][E]	4.13×10^{-3}	0.465	。RT [USER]	7.66×10^{-4}	0.563
NEWS WEB 24	3.70×10^{-3}	1.00	された	6.21×10^{-4}	0.492
。』[URL]	3.68×10^{-3}	0.954	れている	6.17×10^{-4}	0.589
している	3.64×10^{-3}	0.602	してい	5.83×10^{-4}	0.499
のベストセラー→	3.36×10^{-3}	0.984	ではない	3.39×10^{-4}	0.480
番組で紹介	3.22×10^{-3}	0.997	０万円	2.77×10^{-4}	0.83
WEB 24 です	3.21×10^{-3}	1.00	のエネルギー政策	2.26×10^{-4}	0.97
24 です。	3.21×10^{-3}	1.00	、２０	2.25×10^{-4}	0.90
ツイートには	3.18×10^{-3}	1.00	yes or no	2.16×10^{-4}	1.0
で紹介し	3.15×10^{-3}	0.986	• ‘) yes or	1.97×10^{-4}	1.0
してよい	3.14×10^{-3}	0.997	ω • ‘) yes	1.97×10^{-4}	1.0
よいツイートに	3.11×10^{-3}	1.00	? [URL] 拡散	1.97×10^{-4}	1.0
てよいツイート	3.11×10^{-3}	1.00	or no?	1.97×10^{-4}	1.0
編集部）	3.10×10^{-3}	1.00	no? [URL]	1.97×10^{-4}	1.0
SankeiBiz 編集部	3.10×10^{-3}	1.00	、日本の	1.93×10^{-4}	0.82

TABLE IV: Morpheme tri-grams extracted as features representing c_0^-

Ranking by IG			Ranking by BPIG		
Tri-gram f_i	$\mathrm{IG}(C_0\|F_i)$	$p(c_0^+\|f_i)$	Tri-gram f_i	$\mathrm{BPIG}(C_0\|F_i)$	$p(c_0^+\|f_i)$
(笑)	1.84×10^{-3}	0.009	[B][USER] お	1.83×10^{-3}	0.001
[B][USER] お	1.83×10^{-3}	0.001	(笑)	1.60×10^{-3}	0.009
笑)[E]	1.43×10^{-3}	0.005	＼(^o	1.11×10^{-3}	0.006
!! [E]	1.39×10^{-3}	0.044	笑)[E]	1.09×10^{-3}	0.005
＼(^o	1.25×10^{-3}	0.006	[B][USER] おはよう	9.66×10^{-4}	0.002
[B][USER] おはよう	9.66×10^{-4}	0.002	[B][USER] おやすみ	8.98×10^{-4}	0.000
• ω •	9.08×10^{-4}	0.028	[B][USER] そう	8.67×10^{-4}	0.000
[B][USER] おやすみ	8.98×10^{-4}	0.000	´▽‘	6.02×10^{-4}	0.002
[B][USER] そう	8.67×10^{-4}	0.000	[USER] おは	5.98×10^{-4}	0.002
•• [E]	8.24×10^{-4}	0.030	▽‘)	5.89×10^{-4}	0.002
!!!	7.00×10^{-4}	0.061	[B][USER] え	5.76×10^{-4}	0.000

TABLE V: N-grams that frequently co-occur only with a specific hashtag or user in c_0^+ (excerpted)

N-gram f_i	Hashtag or user $\arg\max_{m_{k\in K_i}} \mathrm{IG}(M_k\|F_i)$	Penalty for f_i $\max_{k\in K_i} \mathrm{IG}(M_k\|F_i)$
』[URL][E]	#介護	9.05×10^{-2}
のベストセラー→	#本	4.19×10^{-2}
受験のベストセラー→	#学参	3.99×10^{-2}
SankeiBiz 編集部	#news	3.84×10^{-2}
...[URL][E]	#newsJP	3.60×10^{-2}
NEWS WEB 24	#nhk24	3.48×10^{-2}
番組で紹介し	#nhk24	3.10×10^{-2}
:[URL][E]	@snn007	9.29×10^{-2}
産経新聞）[URL][E]	@selection_news	4.76×10^{-2}
）[URL][E]	@selection_news	3.75×10^{-2}
ヨミドクター）[URL][E]	@yomidr	3.24×10^{-2}

VI. Suggestions, questions, and fact statements with some references (quotation) can be extracted as public concerns from Japanese tweets, according to this analysis result. We suppose that these analyses can be used to define the boundary between positive example c^+ and negative example c^- towards drafting annotation manual and building a concern corpus.

VI. Application

A. Public Debate Using SOCIA Dataset

Citispe@k (pronounced “citi-speak”) is a prototype Web application that supports public debate by utilizing the SOCIA dataset. It provides mobility and reach by supporting Web browsers running on smart phones and tablets. The term citispe@k is based on the idea that citizens speak about social issues and current events of the regions in which they live. Users can discuss and sort out regional issues by referencing news articles, tweets, or other relevant resources on the Web by using citispe@k. By creating discussion topics or inputting opinions into the system, those topics and opinions are also stored as the SOCIA dataset. Figure 7 shows a screenshot of citispe@k. The screenshot has lists of event or related information. Events recently updated are listed on the left of the screenshot. The system initially shows all events. The user can then limit the list to show only events related to a region. When the user selects an event from the list, information about the event is shown on the right side of the screenshot. Information

TABLE VI: Modality types of morpheme N-grams extracted as features representing c_0^+ (excerpted)

Modality	N-gram f_i	BPIG$(C_0\|F_i)$	$p(c_0^+\|f_i)$
Quotation Retweet	」[URL][E]	4.14×10^{-3}	0.758
	ニュース [URL]	1.11×10^{-3}	0.825
	RT [USER]:	7.66×10^{-4}	0.563
	：日本経済新聞	1.46×10^{-4}	0.82
	読売新聞）[URL][E]	1.00×10^{-4}	0.78
	-MSN 産経ニュース	9.19×10^{-5}	0.884
Suggestion Assertion	べき。	2.20×10^{-4}	0.76
	すべき	1.66×10^{-4}	0.725
	べきだ	1.48×10^{-4}	0.61
	するべき	6.03×10^{-5}	0.60
	したらどうだろう	3.92×10^{-5}	1.0
Averment Fact	ている。	1.21×10^{-3}	0.602
	ています	9.14×10^{-4}	0.539
	している	9.14×10^{-4}	0.602
	されている	2.26×10^{-4}	0.629
Question Doubt	•‘) yes or no?	1.97×10^{-4}	1.0
	では？	1.18×10^{-4}	0.63
	ているのか	7.49×10^{-5}	0.55
	のでしょうか	4.19×10^{-5}	0.53
Content	日本の	5.32×10^{-3}	0.824
	億円	1.14×10^{-3}	0.900
	委員会	1.04×10^{-3}	0.884
	兆円	8.79×10^{-4}	0.940
	政策を	3.60×10^{-4}	0.97
	研究機関	2.62×10^{-4}	0.98

consists of news articles, tweets, and events related to the event. Those resources can be easily shown and visualized in an iFrame without leaving the system. Users can append comments, e.g. ideas, questions, and answers, by selecting specific content provided by citispe@k. A comment can also be posted to Twitter (via @citispeak account) to further its reach and be stored in SOCIA. Users can create discussion topics related to events, news articles and tweets. The “View related topics” button (Figure 11) is used to see topics related to the event being viewed. Users can create a new discussion topic about the event by clicking the ”Make a new topic” button. The cycle of the discussions in citispe@k is that users browse events, get topics related to an event, and add their opinion

Citispe@k also has a function supporting concern assessment. The system aim to support the analysis of the trends in citizens’ awareness, its background information, and the anxiety about social issues. For example, a committee for scientific verification of road construction in Aioiyama-

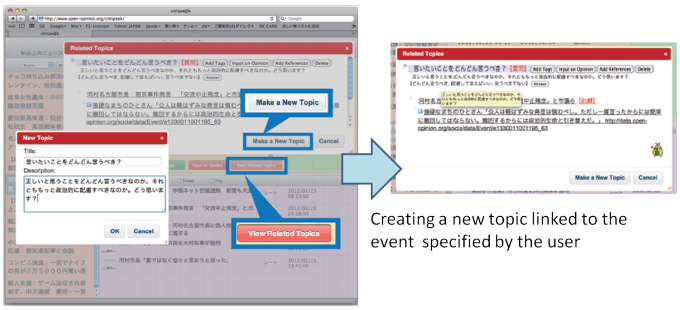


Fig. 11: Creating a new discussion topic

Fig. 12: Annotating selected event with tags representing criteria

Ryokuchi Park in Nagoya City analyzes road construction.[11] A report on their analysis was made based on several criteria: “economic chance”, “life, educational or cultural chance”, “safety, security”, etc. Thus, classifing opinions on the basis of criteria is effective for concern adjustment. Citispe@k provides tags for such criteria. Users can add tags composed of criteria and polarity, such as “Environment +” or “Environment -”. Citispe@k also provides tags that can be used to express the intention of an utterance, like “Question”, “Idea”, and “Refutation”. If events or news articles have many such tags, the tags can be used to support the analysis of concerns. Fig. 12 shows an example of tagging an event. We designed the tags by referencing the QOC model [24] and the Deliberatorium [14] for supporting concern assessment through public debates using citispe@k and the contents in SOCIA.

B. Goal Matching Service Using SOCIA Dataset

We are planning to develop a Web service to match citizens and agents who are aiming at similar goals to facilitate collaboration. Toward this end, we expanded the SOCIA ontology to describe the public goals in Fig. 6. The property `socia:subgoal` enables us to describe the hierarchical structure of goals and subgoals. The public goal matching service that we aim to develop requires high-recall retrieval of similar goals to facilitate inter-domain, inter-area, and inter-organizational collaboration.

Pairs of similar goals are connected by the `schema:isSimilarTo` property[12]. The similarity between public goals can be calculated on the basis of a recursive definition of a bag-of-features vector as:

[11] http://www.city.nagoya.jp/shisei/category/53-3-7-4-0-0-0-0-0-0.html (in Japanese)

[12] http://schema.org/isSimilarTo

$$\mathrm{sim}(g_i, g_j) = \frac{\mathrm{bof}(g_i) \cdot \mathrm{bof}(g_j)}{\|\mathrm{bof}(g_i)\|\|\mathrm{bof}(g_j)\|} \tag{8}$$

$$\mathrm{bof}(g) = \frac{\alpha}{\|\mathrm{tfidf}(g)\|}\mathrm{tfidf}(g) + \frac{\beta}{\|\mathrm{lda}(g)\|}\mathrm{lda}(g) + \frac{\gamma}{|\mathrm{sub}(g)|}\sum_{sg \in \mathrm{sub}(g)} \frac{\mathrm{bof}(sg)}{\|\mathrm{bof}(sg)\|} \tag{9}$$

$$\mathrm{tfidf}(g) = \begin{pmatrix} \mathrm{tfidf}(w_1, g) \\ \vdots \\ \mathrm{tfidf}(w_{|W|}, g) \\ 0 \\ \vdots \\ 0 \end{pmatrix} \in \mathbb{R}^{|W|+|Z|}, \tag{10}$$

$$\mathrm{lda}(g) = \begin{pmatrix} 0 \\ \vdots \\ 0 \\ \mathrm{p}(z_1|g) \\ \vdots \\ \mathrm{p}(z_{|Z|}|g) \end{pmatrix} \in \mathbb{R}^{|W|+|Z|}, \tag{11}$$

where g denotes a public goal, $\mathrm{bof}(g)$ denotes a bag-of-features vector of g, and $\mathrm{sub}(g)$ denotes a set of subgoals of g. Here, $w \in W$ denotes a term, $z \in Z$ denotes a latent topic derived by a latent topic model [25], and $\mathrm{tfidf}(w, g)$ denotes the TF-IDF, i.e., the product of term frequency and inverse document frequency, of w in a title and a description of g. The $\mathrm{p}(z|g)$ denotes the probability of z given g, $0 \leq \alpha, \beta, \gamma \leq 1$, and $\alpha + \beta + \gamma = 1$. The reason this definition incorporates a latent topic model is to enable short descriptions of goals to be dealt with because TF-IDF is insufficient for calculating similarities in short texts. The parameters α, β, and γ are empirically determined on the basis of actual data.

This prototyped method of calculating similarities should be tested, verified, and refined though experiments in future work using the LOD set of public goals that we present.

VII. Conclusion

We designed the SOCIA ontology, which is a linked data model to share context behind local concerns with two approaches: (1) structuring Web news articles and microblogs about local concerns on the basis of geographical regions and events that were referred to by content, and (2) structuring public issues and their solutions as public goals. We moreover built the SOCIA dataset, which was a linked open dataset, on the basis of the SOCIA ontology. Web news articles and microblogs related to local concerns were semi-automatically gathered and structured. 54,854 news articles were stored to the SOCIA dataset and they were automatically linked with prefectures and events. Moreover, 657 public goals were manually extracted from Web content related to revitalization from the Great East Japan Earthquake.

We investigated feature expressions to extract public concerns from microblogs written in Japanese towards more accurate extraction of public concerns. To address a technical issue about sample selection bias in our microblog corpus, we formulated a metric for mining feature expressions, i.e., bias-penalized information gain (BPIG). We conducted an experiment for extracting features representing positive examples and negative examples. The experimental results showed that BPIG is more suitable for dealing with training data with hashtag-dependent sample selection bias than the conventional information gain.

Furthermore, we presented applications of the SOCIA dataset, i.e., a public debate support system and a goal matching service. These applications utilize the SOCIA dataset to share context behind local concerns. We are planning to sophisticate the SOCIA ontology and dataset towards facilitating public collaboration in the real world.

Acknowledgment

This work was supported by a Grant-in-Aid for Young Scientists (B) (No. 25870321) from the Japan Society for the Promotion of Science (JSPS) and SCOPE from the Ministry of Internal Affairs and Communications, Japan.

References

[1] H. Komiyama, "Vision 2050 and the role of Japan toward the sustainable society," in *Proceedings of the 4th International Symposium on Environmentally Conscious Design and Inverse Manufacturing*, 2005, pp. 2–4.

[2] S. Shiramatsu, R. Swezey, H. Sano, N. Hirata, T. Ozono, and T. Shintani, "Structuring Japanese Regional Information Gathered from the Web as Linked Open Data for Use in Concern Assessment," in *Electronic Participation - Proceedings of the 4th IFIP WG 8.5 International Conference, ePart 2012*, ser. Lecture Notes in Computer Science, vol. 7444. Springer, 2012, pp. 73–84.

[3] S. Shiramatsu, N. Hirata, R. Swezey, H. Sano, , T. Ozono, and T. Shintani, "Gathering Public Concerns from Web towards Building Corpus of Japanese Regional Concerns," in *Proceedings of the 2012 IIAI International Conference on Advanced Applied Informatics*, 2012, pp. 248–253.

[4] S. Shiramatsu, T. Ozono, and T. Shintani, "Approaches to Assessing Public Concerns: Building Linked Data for Public Goals and Criteria Extracted from Textual Content," in *Electronic Participation - Proceedings of the 5th IFIP WG 8.5 International Conference, ePart 2013*, ser. Lecture Notes in Computer Science, vol. 8075. Springer, 2013, pp. 109–121.

[5] T. Heath and C. Bizer, *Linked Data: Evolving the Web into a Global Data Space*. Morgan & Claypool, 2011.

[6] J. Hochtl and P. Reichstadter, "Linked open data - a means for public sector information management," in *Proceedings of the 2nd International Conference on Electronic Government and the Information Systems Perspective*, ser. Lecture Notes in Computer Science, vol. 6866. Springer, 2011, pp. 330–343.

[7] International Association for Public Participation, "IAP2 Spectrum of Public Participation," http://www.iap2.org/associations/4748/files/IAP2%20Spectrum_vertical.pdf, 2007.

[8] White House, "Open government initiative," http://www.whitehouse.gov/open, 2009.

[9] N. Benn and A. Macintosh, "Argument visualization for eparticipation: towards a research agenda and prototype tool," in *Electronic Participation - Proceedings of the 3rd IFIP WG 8.5 international conference, ePart 2011*, ser. Lecture Notes in Computer Science, vol. 6847. Springer, 2011, pp. 60–73.

[10] H. Jeong, S. Shiramatsu, T. Hatori, and K. Kobayashi, "Discourse analysis of public debates using corpus linguistic methodologies," *Journal of Computers*, vol. 3, no. 8, pp. 58–68, 2008.

[11] P. Kirschner, S. Shum, and C. Carr, *Visualizing Argumentation: Software Tools for Collaborative and Educational Sense-Making*. Springer, 2003.

[12] A. Selvin and S. Shum, "Hypermedia as a productivity tool for doctoral research," *New Review of Hypermedia and Multimedia, Special Issue on Scholarly Hypermedia*, vol. 11, no. 1, pp. 91–101.

[13] A. D. Liddo and S. B. Shum, "Cohere: A prototype for contested collective intelligence," in *Workshop on Collective Intelligence in Organizations: Toward a Research Agenda, ACM Computer Supported Cooperative Work*, 2010.

[14] L. Iandoli, M. Klein, and G. Zolla, "Enabling online deliberation and collective decision making through large-scale argumentation: A new approach to the design of an internet-based mass collaboration platform," *International Journal of Decision Support System Technology*, vol. 1, no. 1, pp. 69–92, 2009.

[15] C. Reed and G. Rowe, "Araucaria: Software for argument analysis, diagramming and representation," *International Journal of AI Tools*, vol. 13, no. 4, pp. 961–980, 2004.

[16] N. Kamimaeda, N. Izumi, and K. Hasida, "Evaluation of Participants' Contributions in Knowledge Creation Based on Semantic Authoring," *The Learning Organization*, vol. 14, no. 3, pp. 263–280, 2007.

[17] K. Hasida, "Semantic Authoring and Semantic Computing," in *New Frontiers in Artificial Intelligence: Joint Proceeding of the 17th and 18th Annual Conferences of the Japanese Society for Artificial Intelligence*, ser. Lecture Notes in Computer Science, vol. 3609. Springer, 2007, pp. 137–149.

[18] S. W. van den Braak, H. van Oostendorp, H. Prakken, and G. A. W. Vreeswijk, "A critical review of argument visualization tools: Do users become better reasoners?" in *Workshop Notes of the ECAI-2006 Workshop on CMNA*, 2006, pp. 67–75.

[19] A. van Aggelen, "Modelling the european debates," http://www.talkofeurope.eu/2014/05/modelling-the-european-debates/, 2014.

[20] L. Porwol, A. Ojo, and J. Breslin, "A semantic model for e-participation: detailed conceptualization and ontology," in *Proceedings of the 15th Annual International Conference on Digital Government Research*. ACM, 2014, pp. 263–272.

[21] J. Rennie, L. Shih, J. Teevan, and D. Karger, "Tackling the poor assumptions of naive bayes text classifiers," in *Proceedings of the 20th International Conference on Machine Learning*, 2003, pp. 616–623.

[22] R. Swezey, H. Sano, S. Shiramatsu, T. Ozono, and T. Shintani, "Automatic detection of news articles of interest to regional communities," *International Journal of Computer Science and Network Security*, vol. 12, no. 6, pp. 99–106, 2012.

[23] R. Swezey, H. Sano, N. Hirata, S. Shiramatsu, T. Ozono, and T. Shintani, "An e-participation support system for regional communities based on linked open data, classification and clustering," in *Proceedings of the 11th IEEE International Conference on Cognitive Informatics & Cognitive Computing*, 2012, pp. 211–218.

[24] A. MacLean, R. Young, V. Bellotti, and T. Moran, "Questions, options, and criteria: elements of design space analysis," *Human Compututer Interaction*, vol. 6, no. 3, pp. 201–250, 1991.

[25] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent Dirichlet Allocation," *Journal of Machine Learning Research*, vol. 3, pp. 993–1022, 2003.

Timed-Release Certificateless Encryption

Toru Oshikiri
Graduate School of Engineering
Tokyo Denki University
Tokyo, Japan

Taiichi Saito
Tokyo Denki University
Tokyo, Japan

***Abstract*—Timed-Release Encryption(TRE) is an encryption mechanism that allows a receiver to decrypt a ciphertext only after the time that a sender designates. In this paper, we propose the notion of Timed-Release Certificateless Encryption(TRCLE), and define its security models. We also show a generic construction of TRCLE from Public-Key Encryption(PKE), Identity-Based Encryption(IBE) and one-time signature, and prove that the constructed scheme achieves the security we defined.**

***Keywords*—timed-release encryption, identity-based encryption, one-time signature**

I. Introduction

This paper introduces the notion of *Timed-Release Certificateless Encryption (TRCLE)*. TRCLE is a variant of *Timed-Release Encryption (TRE)* [1] [2], in which a sender can generate a ciphertext designating a time to decrypt it, and a receiver can decrypt the ciphertext only after the designated time.

A TRCLE system consists of a key generation center (KGC), a time server (TS), senders and receivers. The KGC helps a receiver to generate a decryption key corresponding to the ID and public key of a receiver. The TS periodically broadcasts a time signal corresponding to the current time. A sender encrypts a message using an ID and a public key of a receiver and a time after which the ciphertext could be decrypted. The receiver decrypts the ciphertext using the decryption key and the time signal corresponding to the time designated by the sender. The TRCLE system does not allow the KGC to obtain the decryption key of receiver, and then it allows only the receiver to decrypt the ciphertext only after the designated time.

The decryption key consists of two keys, a partial secret key and user secret key. Since the former is generated by the KGC and the user but the latter only by the user, the KGC does not know the whole decryption key and cannot decrypt ciphertext.

II. Application

TRCLE has an application to online "sealed-bid auction" in online community in which each registered user has an ID. In the auction system, every user can become auctioneer by publicizing his ID and public key. Each bidder encrypts his price by using the auctioneer's ID and public key and submits the ciphertext as sealed-bid. The auctioneer can decrypt all bids only after the pre-determined closing time. In the sealed-bid auction based on TRCLE, each user determines whether he trusts the auctioneer of ID and attends the auction by checking the reputations and the transaction records in the past auctions organized by the user of ID. Every user easily starts a sealed-bid auction based on TRCLE, since it does not requires heavy infrastructure linking public keys to ID such as Public-Key Infrastructure (PKI).

III. Related Works

There is another variant of TRE, *Timed-Release Identity-Based Encryption (TRIBE)* [3] [4]. In TRIBE, a user can decrypt a ciphertext only when the user has the receiver's secret key and the time signal generated by TS. Then, if the receiver does not have the time signal or the TS does not have the secret key, they cannot decrypt the ciphertext. In [3], two security models of TRIBE are defined. One is security against malicious receiver, $\mathsf{IND\text{-}ID\text{-}CCA}_{\mathsf{CR}}$ security. The other is security against malicious TS, $\mathsf{IND\text{-}ID\text{-}CCA}_{\mathsf{TS}}$ security. A generic construction of TRIBE that achieves the security is also shown in [3]. It is a combination of two IBE schemes and a one-time signature schemes, based on "Parallel Encryption" by Dodis-Katz [5], and the security is proved in the standard model.

TRCLE has an advantage over TRIBE in that a compromised KGC cannot decrypt any ciphertext since the key generation process is split between the KGC and the user. Then we discuss only security against malicious KGC in this paper. The other security is proved in almost the same way as in TRIBE.

TRCLE can be considered also a variant of *Certificateless Encryption (CLE)* [6] having the mechanism of TRE. In CLE, the decryption key is partially determined by KGC

IV. Contributions

In this paper, we introduce timed-release certificateless encryption (TRCLE) and define its security models including security against malicious KGC, $\mathsf{Mal.KGC}$ security. We also present a generic construction of TRCLE. It is a combination of a Public-Key Encryption(PKE) scheme, two Identity-Based Encryption(IBE) schemes and a one-time signature schemes, also based on "Parallel Encryption". We see that if the primitive PKE scheme is *indistinguishability against adaptive chosen ciphertext attacks*($\mathsf{IND\text{-}CCA}$) secure and the primitive one-time signature scheme is *one-time strong existential unforgeability against chosen message attacks*($\mathsf{OT\text{-}sEUF\text{-}CMA}$) secure, then the constructed TRCLE scheme is $\mathsf{Mal.KGC}$ secure in the standard model.

V. PRELIMINARIES

In this section, we review public-key encryption (PKE), identity-based encryption (IBE) and one-time signature, which we use later.

A. Public-Key Encryption

Let λ be a security parameter. An *public-key encryption scheme* $\mathcal{PKE}$ [7] consists of three probabilistic polynomial-time algorithms $\mathcal{PKE}$ = (PKE.Gen, PKE.Enc, PKE.Dec). The key generation algorithm PKE.Gen takes λ as input, and outputs a public key pk and a secret key sk. The encryption algorithm PKE.Enc takes pk, a message m as inputs, and outputs a ciphertext c. The decryption algorithm PKE.Dec takes a secret key sk and a ciphertext c as inputs, and outputs the plaintext m' or $\perp$. These algorithms are assumed to satisfy that if (pk, sk) = $\mathsf{PKE.Gen}(\lambda)$ then $\mathsf{PKE.Dec}(sk, \mathsf{PKE.Enc}(pk, m)) = m$ for any m.

1) IND-CCA *Security:* We review a standard security notion for PKE: *indistinguishability against adaptive chosen ciphertext attacks (*IND-CCA*)* security. We here describe the IND-CCA security for PKE scheme $\mathcal{PKE}$ based on the following IND-CCA game between a challenger $\mathcal{C}$ and an adversary $\mathcal{A}$.

Setup
$\mathcal{C}$ runs $(pk, sk) \leftarrow \mathsf{PKE.Gen}(\lambda)$. $\mathcal{C}$ sends pk to $\mathcal{A}$ and keeps sk secret.

Phase1
$\mathcal{A}$ can adaptively issue *decryption queries* c. $\mathcal{C}$ responds to a decryption query c by running $m' = \mathsf{PKE.Dec}(sk, c)$, and returning m' to $\mathcal{A}$.

Challenge
$\mathcal{A}$ sends two messages m_0, m_1 such that $|m_0| = |m_1|$ to $\mathcal{C}$. $\mathcal{C}$ randomly chooses $b \in \{0, 1\}$ and sends a challenge ciphertext $c^* = \mathsf{PKE.Enc}(pk, m_b)$ to $\mathcal{A}$.

Phase2
$\mathcal{A}$ can adaptively issue *decryption queries* c in the same way as in **Phase1** except that the decryption queries c must differ from the challenge ciphertext c^*.

Guess
$\mathcal{A}$ outputs a guess $b' \in \{0, 1\}$ and wins if $b = b'$.

We define an advantage of $\mathcal{A}$ in the IND-CCA game as $Adv^{\mathsf{IND\text{-}CCA}}_{\mathcal{PKE},\mathcal{A}}(\lambda) = |\Pr[b = b'] - \frac{1}{2}|$, in which the probability is taken over the random coins used by $\mathcal{C}$ and $\mathcal{A}$. We say that the PKE scheme $\mathcal{PKE}$ is IND-CCA *secure* if, for any probabilistic polynomial-time adversary $\mathcal{A}$, the function $Adv^{\mathsf{IND\text{-}CCA}}_{\mathcal{PKE},\mathcal{A}}(\lambda)$ is negligible in λ.

B. Identity-Based Encryption

Let λ be a security parameter. An *identity-based encryption scheme* $\mathcal{IBE}$ [8] consists of four probabilistic polynomial-time algorithms $\mathcal{IBE}$ = (IBE.Setup, IBE.Ext, IBE.Enc, IBE.Dec). The setup algorithm IBE.Setup takes λ as input, and outputs a public parameter $params$ and a master secret key msk. The extract algorithm IBE.Ext takes $params$, msk, and an identity ID as inputs, and outputs a decryption key d_{ID}. The encryption algorithm IBE.Enc takes $params, \mathsf{ID}$, a message m as inputs, and outputs a ciphertext c. The decryption algorithm IBE.Dec takes $params$, a decryption key d_{ID} and a ciphertext c as inputs, and outputs the plaintext m' or $\perp$. These algorithms are assumed to satisfy that if $(params, msk) = \mathsf{IBE.Setup}(\lambda)$ and $d_{\mathsf{ID}} = \mathsf{IBE.Ext}(params, msk, \mathsf{ID})$ then $\mathsf{IBE.Dec}(params, d_{\mathsf{ID}}, \mathsf{IBE.Enc}(params, \mathsf{ID}, m)) = m$ for any m.

1) IND-ID-CCA *Security:* We review a standard security notion for IBE: *indistinguishability against adaptive identity and chosen ciphertext attacks (*IND-ID-CCA*)* security [9]. We here describe the IND-ID-CCA security for IBE scheme $\mathcal{IBE}$ based on the following IND-ID-CCA game between a challenger $\mathcal{C}$ and an adversary $\mathcal{A}$.

Setup
$\mathcal{C}$ runs $(params, msk) \leftarrow \mathsf{IBE.Setup}(\lambda)$. $\mathcal{C}$ sends $params$ to $\mathcal{A}$ and keeps msk secret.

Phase1
$\mathcal{A}$ can adaptively issue *extraction queries* ID and *decryption queries* (ID, c). $\mathcal{C}$ responds to an extraction query ID by running $d_{\mathsf{ID}} = \mathsf{IBE.Ext}(params, msk, \mathsf{ID})$ and returning d_{ID} to $\mathcal{A}$. $\mathcal{C}$ responds to a decryption query (ID, c) by running $d_{\mathsf{ID}} = \mathsf{IBE.Ext}(params, msk, \mathsf{ID})$ and $m' = \mathsf{IBE.Dec}(params, d_{\mathsf{ID}}, c)$, and returning m' to $\mathcal{A}$.

Challenge
$\mathcal{A}$ sends two messages m_0, m_1 such that $|m_0| = |m_1|$, and an identity to be challenged ID^* to $\mathcal{C}$. The challenge identity ID^* must differ from any ID issued as extraction query in **Phase1**. $\mathcal{C}$ randomly chooses $b \in \{0, 1\}$ and sends a challenge ciphertext $c^* = \mathsf{IBE.Enc}(params, \mathsf{ID}^*, m_b)$ to $\mathcal{A}$.

Phase2
$\mathcal{A}$ can adaptively issue extraction queries ID and decryption queries (ID, c) in the same way as in **Phase1** except that the extraction queries ID must differ from the challenge identity ID^*, and decryption queries (ID, c) must differ from the pair (ID^*, c^*).

Guess
$\mathcal{A}$ outputs a guess $b' \in \{0, 1\}$ and wins if $b = b'$.

We define an advantage of $\mathcal{A}$ in the IND-ID-CCA game as $Adv^{\mathsf{IND\text{-}ID\text{-}CCA}}_{\mathcal{IBE},\mathcal{A}}(\lambda) = |\Pr[b = b'] - \frac{1}{2}|$, in which the probability is taken over the random coins used by $\mathcal{C}$ and $\mathcal{A}$. We say that the IBE scheme $\mathcal{IBE}$ is IND-ID-CCA *secure* if, for any probabilistic polynomial-time adversary $\mathcal{A}$, the function $Adv^{\mathsf{IND\text{-}ID\text{-}CCA}}_{\mathcal{IBE},\mathcal{A}}(\lambda)$ is negligible in λ.

C. One-time Signature

Let λ be a security parameter. $\mathcal{A}$ *signature* scheme $\mathcal{SIG}$ consists of three probabilistic polynomial-time algorithms $\mathcal{SIG} = (\mathsf{SigGen}, \mathsf{Sign}, \mathsf{Verify})$. The key generation algorithm SigGen takes λ as input, and outputs a signing key sk and a verification key vk. The signing algorithm Sign takes sk and a message m as inputs , and outputs a signature σ. The verification algorithm Verify takes vk, a message m, and a signature σ as inputs, and outputs `accept` or `reject`. These

algorithms are assumed to satisfy that if $(sk, vk) = \mathsf{SigGen}(\lambda)$ then $\mathsf{Verify}(vk, m, \mathsf{Sign}(sk, m)) = \mathtt{accept}$ for any m.

1) OT-sEUF-CMA *Security:* We review a security notion for one-time signature scheme: *one-time strong existential unforgeability against chosen message attacks (*OT-sEUF-CMA*)* security [10]. We here describe the OT-sEUF-CMA security for signature scheme $\mathcal{SIG}$ based on the following OT-sEUF-CMA game between a challenger $\mathcal{C}$ and an adversary $\mathcal{A}$.

Setup

$\mathcal{C}$ runs the (*sk, vk*) $\leftarrow \mathsf{SigGen}(\lambda)$. $\mathcal{C}$ sends vk to $\mathcal{A}$ and keeps sk secret.

Query

$\mathcal{A}$ can issue a *signing query* m to $\mathcal{C}$ only once. $\mathcal{C}$ responds to the singing query m by running $\sigma = \mathsf{Sign}(vk, m)$ and returning σ to $\mathcal{A}$.

Forge

$\mathcal{A}$ outputs a pair (m^*, σ^*).

We define the advantage of $\mathcal{A}$ in the OT-sEUF-CMA game as $Adv^{\text{OT-sEUF-CMA}}_{\mathcal{SIG},\mathcal{A}}(\lambda) = \Pr[\mathsf{Verify}(vk, m^*, \sigma^*) = \mathtt{accept} \wedge (m, \sigma) \neq (m^*, \sigma^*)]$, in which the probability is taken over the random coins used by $\mathcal{C}$ and $\mathcal{A}$. We say that the signature scheme $\mathcal{SIG}$ is OT-sEUF-CMA *secure* if, for any probabilistic polynomial-time adversary $\mathcal{A}$, the function $Adv^{\text{OT-sEUF-CMA}}_{\mathcal{SIG},\mathcal{A}}(\lambda)$ is negligible in λ.

VI. Timed-Release Certificateless Encryption(TRCLE)

In this section, we introduce timed-release certificateless encryption(TRCLE) scheme and define its security models.

A TRCLE system consists of a key generation center (KGC), a time server (TS), senders and receivers. The KGC helps a receiver to generate a partial secret key corresponding to an ID of the receiver. The TS periodically broadcasts a time signal corresponding to the current time. A sender encrypts a message using the ID and public key of a receiver, and a time after which the ciphertext could be decrypted. The receiver decrypts the ciphertext using the partial secret key, the user secret key and the time signal corresponding to the time designated by the sender.

Let λ be a security parameter. An *timed-release certificateless encryption scheme* $\mathcal{TRCLE}$ consists of seven probabilistic polynomial-time algorithms $\mathcal{TRCLE}$ = (KGC_Setup, TS_Setup, PartialKeyGen, UserKeyGen, Release, Encrypt, Decrypt). The key generation center's setup algorithm KGC_Setup takes λ as input, and outputs a public parameter $params$ and a master secret key msk. The time server's setup algorithm TS_Setup takes λ as input, and outputs a public key tpk and the corresponding secret key tsk. The partial secret key generation algorithm PartialKeyGen takes $params, msk$ and ID as input, and outputs a partial secret key psk_{ID} corresponding to ID. The user key generation algorithm UserKeyGen takes $params$ and ID as input, and outputs a user public key upk_{ID} and a user secret key usk_{ID} corresponding to ID. The release algorithm Release takes tpk, tsk and a time period T as inputs, and outputs a time signal d_T. The encryption algorithm Encrypt takes $params$, tpk, ID, upk_{ID}, T and a message m as inputs, and outputs a ciphertext c. The decryption algorithm Decrypt takes as inputs $params$, tpk, psk_{ID}, usk_{ID}, d_T and a ciphertext c', and outputs the plaintext m' or $\perp$. These algorithms are assumed to satisfy that Decrypt($params$, tpk, psk_{ID}, usk_{ID}, d_T, Encrypt($params$, tpk, ID, upk_{ID}, T, m)) = m holds for any m, if (tpk, tsk) = TS_Setup(λ), ($params$, msk) = KGC_Setup(λ), psk_{ID} = PartialKeyGen($params, msk,$ ID), ($upk_{\mathsf{ID}}, usk_{\mathsf{ID}}$) = UserKeyGen($params,$ ID) and d_T = TR.Release(tpk, tsk, T) hold.

A. Security

We can consider security against KGC, TS, receiver or outsider. Since the TS security is implied by KGC security, we present the three kinds of security.

1) Mal.KGC *Security:* We introduce a security notion for TRCLE: *indistinguishability against adaptive identity and chosen ciphertext attacks by key generation center (*Mal.KGC*) security*. This security ensures that a malicious key generation center, who has a master secret key msk, cannot obtain any information of message from a ciphertext without a user secret key usk_{ID}. We here describe the Mal.KGC security for a TRCLE scheme $\mathcal{TRCLE}$ based on the following Mal.KGC game between a challenger $\mathcal{C}$ and adversary $\mathcal{A}$.

Setup

$\mathcal{C}$ runs (*tpk, tsk*) $\leftarrow$ TS_Setup(λ) and sends (λ, tpk, tsk) to $\mathcal{A}$. $\mathcal{A}$ runs ($params, msk$) $\leftarrow$ KGC_Setup(λ) and sends $params$ to $\mathcal{C}$. $\mathcal{C}$ creates an empty list List.

Phase1

$\mathcal{A}$ can adaptively issue the following four queries.

Create User query

$\mathcal{A}$ sends (ID, psk_{ID}) to $\mathcal{C}$. When ID is in List, $\mathcal{C}$ returns upk_{ID} corresponding to ID. When ID is not in List, $\mathcal{C}$ runs ($upk_{\mathsf{ID}}, usk_{\mathsf{ID}}$) $\leftarrow$ UserKeyGen($params$, ID) and stores (ID, $psk_{\mathsf{ID}}, upk_{\mathsf{ID}}, usk_{\mathsf{ID}}$) in List. $\mathcal{C}$ returns upk_{ID} corresponding to ID.

Reveal Secret Key query

$\mathcal{A}$ sends ID to $\mathcal{C}$. When ID is in List, $\mathcal{C}$ returns usk_{ID} corresponding to ID. When ID is not in List, $\mathcal{C}$ returns $\perp$

Replace query

$\mathcal{A}$ sends (ID, upk', usk') to $\mathcal{C}$. When ID is in List, $\mathcal{C}$ replaces (ID, $psk_{\mathsf{ID}}, upk_{\mathsf{ID}}, usk_{\mathsf{ID}}$) with (ID, $psk_{\mathsf{ID}}, upk', usk'$). If $usk' = \perp$, $\mathcal{C}$ sets $usk' = usk_{\mathsf{ID}}$ When ID is not in List, $\mathcal{C}$ do nothing.

Decrypt query

$\mathcal{A}$ sends (ID, T, c) to $\mathcal{C}$. When ID is in List, $\mathcal{C}$ runs $d_T \leftarrow \mathsf{Release}(tpk, tsk, T)$ and $m \leftarrow \mathsf{Decrypt}(params, tpk, psk_{\mathsf{ID}}, usk_{\mathsf{ID}}, d_T, c)$ and returns m. When ID is not in List, $\mathcal{C}$ returns $\perp$.

Challenge

$\mathcal{A}$ sends two messages m_0, m_1 such that $|m_0| = |m_1|$, an identity to be challenged ID* and a time period T^* to $\mathcal{C}$. The challenge identity ID* must differ from any ID issued as Replace

queries in **Phase1**. $\mathcal{C}$ randomly chooses $b \in \{0,1\}$ and sends a challenge ciphertext $c^* = \mathsf{Encrypt}(params, tpk, \mathsf{ID}^*, upk_{\mathsf{ID}^*}, T^*, m_b)$ to $\mathcal{A}$.

Phase2

$\mathcal{A}$ can adaptively issue the above four queries in the same way as **Phase1** except that the Replace queries ID must differ from the challenge identity ID^*, and the decryption queries (ID, T, c) must differ from the tuple $(\mathsf{ID}^*, T^*, c^*)$.

Guess

$\mathcal{A}$ outputs a guess $b' \in \{0,1\}$ and wins if $b = b'$.

We define an advantage of $\mathcal{A}$ in the Mal.KGC game as $Adv^{\mathsf{Mal.KGC}}_{\mathcal{TRCLE},\mathcal{A}}(\lambda) = |\Pr[b = b'] - \frac{1}{2}|$, in which the probability is taken over the random coins used by $\mathcal{C}$ and $\mathcal{A}$. We say that the TRCLE scheme $\mathcal{TRCLE}$ is Mal.KGC *secure* if, for any probabilistic polynomial-time adversary $\mathcal{A}$, the function $Adv^{\mathsf{Mal.KGC}}_{\mathcal{TRCLE},\mathcal{A}}(\lambda)$ is negligible in λ.

2) Mal.Receiver *Security:* We introduce a security notion for TRCLE: *indistinguishability against adaptive identity and chosen ciphertext attacks by receiver (*Mal.Receiver*) security*. This security ensures that a malicious receiver, who has a partial secret key psk_{ID} and user secret key usk_{ID}, cannot obtain any information of message from a ciphertext without a time signal d_T corresponding to the time designated by the sender. We here describe the Mal.Receiver security for a TRCLE scheme $\mathcal{TRCLE}$ based on the following Mal.Receiver game between a challenger $\mathcal{C}$ and adversary $\mathcal{A}$.

Setup

$\mathcal{C}$ runs (*params, msk*) $\leftarrow \mathsf{KGC\_Setup}(\lambda)$, (*tpk, tsk*) $\leftarrow \mathsf{TS\_Setup}(\lambda)$, and sends (λ, tpk, tsk) to $\mathcal{A}$. $\mathcal{C}$ creates an empty list `List`.

Phase1

$\mathcal{A}$ can adaptively issue the following five queries.

Create User query

$\mathcal{A}$ sends (ID, upk_{ID}, usk_{ID}) to $\mathcal{C}$. When ID is in `List`, $\mathcal{C}$ do nothing. When ID is not in `List`, $\mathcal{C}$ runs $psk_{\mathsf{ID}} \leftarrow \mathsf{PartialKeyGen}(params, msk, \mathsf{ID})$ and stores $(\mathsf{ID}, psk_{\mathsf{ID}}, upk_{\mathsf{ID}}, usk_{\mathsf{ID}})$ in `List`.

Reveal Partial Key query

$\mathcal{A}$ sends ID to $\mathcal{C}$. When ID is in `List`, $\mathcal{C}$ returns psk_{ID} corresponding to ID. When ID is not in `List`, $\mathcal{C}$ returns $\perp$

Replace query

$\mathcal{A}$ sends $(\mathsf{ID}, upk', usk')$ to $\mathcal{C}$. When ID is in `List`, $\mathcal{C}$ replaces $(\mathsf{ID}, psk_{\mathsf{ID}}, upk_{\mathsf{ID}}, usk_{\mathsf{ID}})$ with $(\mathsf{ID}, psk_{\mathsf{ID}}, upk', usk')$. If $usk' = \perp$, $\mathcal{C}$ sets $usk' = usk_{\mathsf{ID}}$ When ID is not in `List`, $\mathcal{C}$ do nothing.

Release query

$\mathcal{A}$ sends T to $\mathcal{C}$. $\mathcal{C}$ runs $d_T \leftarrow \mathsf{Release}(tpk, tsk, T)$ and returns d_T.

Decrypt query

$\mathcal{A}$ sends (ID, T, c) to $\mathcal{C}$. When ID is in `List`, $\mathcal{C}$ runs $d_T \leftarrow \mathsf{Release}(tpk, tsk, T)$ and $m \leftarrow \mathsf{Decrypt}(params, tpk, psk_{\mathsf{ID}}, usk_{\mathsf{ID}}, d_T, c)$ and returns m. When ID is not in `List`, $\mathcal{C}$ returns $\perp$.

Challenge

$\mathcal{A}$ sends two messages m_0, m_1 such that $|m_0| = |m_1|$, an identity to be challenged ID^* and a time period T^* to $\mathcal{C}$. The time period T^* must differ from any T issued as Release queries in **Phase1**. $\mathcal{C}$ randomly chooses $b \in \{0,1\}$ and sends a challenge ciphertext $c^* = \mathsf{Encrypt}(params, tpk, \mathsf{ID}^*, upk_{\mathsf{ID}^*}, T^*, m_b)$ to $\mathcal{A}$.

Phase2

$\mathcal{A}$ can adaptively issue the above five queries in the same way as **Phase1** except that the Release queries T must differ from T^*, and the decryption queries (ID, T, c) must differ from the tuple $(\mathsf{ID}^*, T^*, c^*)$.

Guess

$\mathcal{A}$ outputs a guess $b' \in \{0,1\}$ and wins if $b = b'$.

We define an advantage of $\mathcal{A}$ in the Mal.Receiver game as $Adv^{\mathsf{Mal.Receiver}}_{\mathcal{TRCLE},\mathcal{A}}(\lambda) = |\Pr[b = b'] - \frac{1}{2}|$, in which the probability is taken over the random coins used by $\mathcal{C}$ and $\mathcal{A}$. We say that the TRCLE scheme $\mathcal{TRCLE}$ is Mal.Receiver *secure* if, for any probabilistic polynomial-time adversary $\mathcal{A}$, the function $Adv^{\mathsf{Mal.Receiver}}_{\mathcal{TRCLE},\mathcal{A}}(\lambda)$ is negligible in λ.

3) Outsider *Security:* We introduce a security notion for TRCLE: *indistinguishability against adaptive identity and chosen ciphertext attacks by outsider (*Outsider*) security*. This security ensures that a outsider, who has a public parameter $params$ and tpk, cannot obtain any information of message from a ciphertext without user secret key usk_{ID}. We here describe the Outsider security for a TRCLE scheme $\mathcal{TRCLE}$ based on the following Outsider game between a challenger $\mathcal{C}$ and adversary $\mathcal{A}$.

Setup

$\mathcal{C}$ runs (*params, msk*) $\leftarrow \mathsf{KGC\_Setup}(\lambda)$, (*tpk, tsk*) $\leftarrow \mathsf{TS\_Setup}(\lambda)$, and sends $(params, tpk)$ to $\mathcal{A}$. $\mathcal{C}$ creates an empty list `List`.

Phase1

$\mathcal{A}$ can adaptively issue the following six queries.

Create User query

$\mathcal{A}$ sends ID to $\mathcal{C}$. When ID is in `List`, $\mathcal{C}$ returns upk_{ID} corresponding to ID. When ID is not in `List`, $\mathcal{C}$ runs $psk_{\mathsf{ID}} \leftarrow \mathsf{PartialKeyGen}(params, msk, \mathsf{ID})$ and $(upk_{\mathsf{ID}}, usk_{\mathsf{ID}}) \leftarrow \mathsf{UserKeyGen}(params, \mathsf{ID})$, and stores $(\mathsf{ID}, psk_{\mathsf{ID}}, upk_{\mathsf{ID}}, usk_{\mathsf{ID}})$ in `List`. $\mathcal{C}$ returns upk_{ID} corresponding to ID.

Reveal Partial Key query

$\mathcal{A}$ sends ID to $\mathcal{C}$. When ID is in `List`, $\mathcal{C}$ returns psk_{ID} corresponding to ID. When ID is not in `List`, $\mathcal{C}$ returns $\perp$

Reveal Secret Key query

$\mathcal{A}$ sends ID to $\mathcal{C}$. When ID is in `List`, $\mathcal{C}$ returns usk_{ID} corresponding to ID. When ID is not in `List`, $\mathcal{C}$ returns $\perp$

Replace query

$\mathcal{A}$ sends $(\mathsf{ID}, upk', usk')$ to $\mathcal{C}$. When ID is in `List`, $\mathcal{C}$ replaces $(\mathsf{ID}, psk_{\mathsf{ID}}, upk_{\mathsf{ID}}, usk_{\mathsf{ID}})$ with $(\mathsf{ID}, psk_{\mathsf{ID}}, upk', usk')$. If $usk' = \perp$, $\mathcal{C}$ sets $usk' = usk_{\mathsf{ID}}$ When ID is not in `List`, $\mathcal{C}$ do nothing.

Release query

$\mathcal{A}$ sends T to $\mathcal{C}$. $\mathcal{C}$ runs $d_T \leftarrow$ Release(tpk, tsk, T) and returns d_T.

Decrypt query

$\mathcal{A}$ sends (ID, T, c) to $\mathcal{C}$. When ID is in List, $\mathcal{C}$ runs $d_T \leftarrow$ Release(tpk, tsk, T) and $m \leftarrow$ Decrypt($params, tpk, psk_{\mathsf{ID}}, usk_{\mathsf{ID}}, d_T, c$) and returns m. When ID is not in List, $\mathcal{C}$ returns $\perp$.

Challenge

$\mathcal{A}$ sends two messages m_0, m_1 such that $|m_0| = |m_1|$, an identity to be challenged ID^* and a time period T^* to $\mathcal{C}$. The challenge identity ID^* must differ from any ID issued as Reveal Partial Key queries in **Phase1**. $\mathcal{C}$ randomly chooses $b \in \{0,1\}$ and sends a challenge ciphertext $c^* =$ Encrypt($params, tpk, \mathsf{ID}^*, upk_{\mathsf{ID}^*}, T^*, m_b$) to $\mathcal{A}$.

Phase2

$\mathcal{A}$ can adaptively issue the above six queries in the same way as **Phase1** except that the Reveal Secret Key queries ID must differ from ID^*, and the decryption queries (ID, T, c) must differ from the tuple (ID^*, T^*, c^*).

Guess

$\mathcal{A}$ outputs a guess $b' \in \{0,1\}$ and wins if $b = b'$.

We define an advantage of $\mathcal{A}$ in the Outsider game as $Adv^{\mathsf{Outsider}}_{\mathcal{TRCLE},\mathcal{A}}(\lambda) = |\Pr[b = b'] - \frac{1}{2}|$, in which the probability is taken over the random coins used by $\mathcal{C}$ and $\mathcal{A}$. We say that the TRCLE scheme $\mathcal{TRCLE}$ is Outsider *secure* if, for any probabilistic polynomial-time adversary $\mathcal{A}$, the function $Adv^{\mathsf{Outsider}}_{\mathcal{TRCLE},\mathcal{A}}(\lambda)$ is negligible in λ.

VII. Construction

Here we present a generic construction of TRCLE scheme from PKE scheme, IBE scheme and one-time signature scheme.

A. Construction

Let Δ = (PKE.Gen, PKE.Enc, PKE.Dec) be a public-key encryption scheme, Π = (IBE.Setup, IBE.Ext, IBE.Enc, IBE.Dec) and Π' = (IBE′.Setup, IBE′.Ext, IBE′.Enc, IBE′.Dec) be identity-based encryption schemes, and Σ = (SigGen, Sign, Verify) be a one-time signature scheme.

A TRCLE scheme Γ = (KGC_Setup, TS_Setup, PartialKeyGen, UserKeyGen, Release, Encrypt, Decrypt) is constructed as follows.

KGC_Setup(λ)**:**
Step 1: Run IBE.Setup on input λ to generate ($params$, msk).
Step 2: Return ($params, msk$).

TS_Setup(λ)**:**
Step 1: Run IBE′.Setup on input λ to generate ($params'$, msk').
Step 2: Set $tpk = params'$ and $tsk = msk'$.
Step 3: Return (tpk, tsk).

PartialKeyGen($params, msk, ID$)**:**
Step 1: Run IBE.Ext($params, msk, ID$) to obtain d_{ID}.
Step 2: Return d_{ID}.

UserKeyGen($params, ID$)**:**
Step 1: Run PKE.Gen on input λ to generate (pk, sk).
Step 2: Set $upk_{ID} = pk$ and $usk_{\mathsf{ID}} = sk$.
Step 3: Return (upk_{ID}, usk_{ID}).

Release(tpk, tsk, T)**:**
Step 1: Run IBE′.Ext(tpk, tsk, T) to obtain d_T.
Step 2: Return d_T.

Encrypt($params, tpk, ID, upk_{ID}, T, m$)**:**
Step 1: Run SigGen on input λ to generate (sk, vk).
Step 2: Randomly choose $s_1 \in \{0,1\}^{|m|}$ and $s_2 \in \{0,1\}^{|m|}$.
Step 3: Compute $s_3 = s_1 \oplus s_2 \oplus m$.
Step 4: Compute c_1=IBE.Enc($params$, ID, $s_1||vk$).
Step 5: Compute c_2=IBE′.Enc(tpk, T, $s_2||vk$).
Step 6: Compute c_3=PKE.Enc($upk_{ID}, s_3||vk$).
Step 7: Compute $\sigma = \mathsf{Sign}(sk, c_1||c_2||c_3||ID||T)$.
Step 8: Set $c = (c_1, c_2, c_3, ID, T, vk, \sigma)$.
Step 9: Return c.

Decrypt($params, tpk, psk_{ID}, usk_{ID}, d_T, c$)**:**
Step 1: Parse c as $c = (c_1, c_2, c_3, ID, T, vk, \sigma)$.
Step 2: If Verify($vk, c_1||c_2||c_3||ID||T, \sigma$) = `reject` , then return $\perp$ and stop.
Step 3: Compute $s_1||vk'$ = IBE.Dec($params$, psk_{ID}, c_1).
Step 4: Compute $s_2||vk''$ = IBE′.Dec(tpk, d_T, c_2).
Step 5: Compute $s_3||vk'''$ = PKE.Dec(usk_{ID}, c_3).
Step 6: If $vk = vk' = vk'' = vk'''$, then return $m = s_1 \oplus s_2 \oplus s_3$ else return $\perp$.

B. Security of TRCLE

1) Mal.KGC *security:*

Theorem 1: If Δ is an IND-CCA secure public-key encryption scheme and Σ is an OT-sEUF-CMA secure one-time signature scheme, then Γ is a Mal.KGC secure timed-release certificateless encryption scheme.

Proof(Theorem 1) Suppose $\mathcal{A}$ is an adversary that breaks the Mal.KGC security of Γ. We construct a simulator $\mathcal{B}$ which breaks the IND-CCA security of the PKE scheme Δ using $\mathcal{A}$. Here we say a ciphertext $c = (c_1, c_2, c_3, \mathsf{ID}, T, vk, \sigma)$ is $valid$ if Verify($vk, c_1||c_2||c_3||\mathsf{ID}||T, \sigma$) = `accept`. Let $c^* = (c_1^*, c_2^*, c_3^*, \mathsf{ID}^*, T^*, vk^*, \sigma^*)$ be the challenge ciphertext. Let `Forge` denote the event that $\mathcal{A}$ submits a valid ciphertext $c = (c_1, c_2, c_3, T, \mathsf{ID}, vk^*, \sigma)$ as a Decrypt query to $\mathcal{C}$ in the **Phase2**, and `Succ` denote the event that $\mathcal{A}$ wins the Mal.KGC game. We assume that $\mathcal{A}$ issues at most q distinct Create User queries. We prove the following claims.

Claim 1: $\Pr[\texttt{Forge}]$ is negligible.

Claim 2: $|\Pr[\texttt{Succ} \wedge \overline{\texttt{Forge}}] + \frac{1}{2}\Pr[\texttt{Forge}] - \frac{1}{2}|$ is negligible.

Proof(Claim 1) We assume `Forge` occurs. Then, we construct a forger $\mathcal{F}$ who breaks OT-sEUF-CMA security of the one-time signature scheme Σ, from $\mathcal{A}$. The description of $\mathcal{F}$ is as follows.

Setup

$\mathcal{F}$ receives vk^* from $\mathcal{C}$. $\mathcal{F}$ runs (tpk, tsk) $\leftarrow$ TS_Setup(λ) and sends (λ, tpk, tsk) to $\mathcal{A}$.

Query

$\mathcal{F}$ responds $\mathcal{A}$'s four queries as the challenger in the Mal.KGC game. If $\mathcal{A}$ happens to issue a valid ciphertext $c = (c_1, c_2, c_3, \mathsf{ID}, T, vk^*, \sigma)$

as decryption query to $\mathcal{F}$ before **Challenge** in the Mal.KGC game, then $\mathcal{F}$ simply outputs $(c_1||c_2||c_3||T||\mathsf{ID}, \sigma)$ as forgery and stops.

Challenge

If $\mathcal{A}$ outputs (m_0, m_1, ID^*, T^*) as challenge , $\mathcal{F}$ randomly chooses $s_1 \in \{0,1\}^{|m|}$, $s_2 \in \{0,1\}^{|m|}$ and $b \in \{0,1\}$, and computes $s_3 = s_1 \oplus s_2 \oplus m_b$. Then $\mathcal{F}$ computes c_1^* = IBE.Enc($params$, ID^*, $s_1||vk^*$), c_2^* = IBE′.Enc(tpk, T^*, $s_2||vk^*$) and c_3^* = PKE.Enc(pub_{ID^*}, $s_3||vk^*$), then issues $m^* = (c_1||c_2||c_3||\mathsf{ID}^*||T^*)$ as signing query to $\mathcal{C}$ and obtains σ^*. Finally $\mathcal{F}$ returns $c^* = (c_1^*, c_2^*, c_3^*, \mathsf{ID}^*, T^*, vk^*, \sigma^*)$ as the challenge ciphertext to $\mathcal{A}$.

Forge

If $\mathcal{A}$ issues a valid ciphertxt c = (c_1, c_2, c_3, ID, T, vk^*, σ) as decryption query, then $\mathcal{F}$ outputs $(c_1||c_2||c_3||\mathsf{ID}||T, \sigma)$ as forgery.

$\mathcal{F}$ can forge the signature if $\mathcal{A}$ issues a decryption query that causes the event `Forge`. It, however, contradicts that Σ is OT-sEUF-CMA secure. Thus, $\Pr[\texttt{Forge}]$ is negligible. □

Proof(Claim 2) We construct an adversary $\mathcal{B}$ who breaks IND-CCA security of the PKE scheme Δ using $\mathcal{A}$. The description of $\mathcal{B}$ is as follows.

Setup

$\mathcal{B}$ receives pk^* from $\mathcal{C}$. Then $\mathcal{B}$ runs (tpk, tsk) ← TS_Setup(λ) and sends (λ, tpk, tsk) to $\mathcal{A}$ and randomly chooses index $i \in \{1, 2, \cdots q\}$. $\mathcal{B}$ creates an empty list `List`.

Phase1

The response of $\mathcal{B}$ for $\mathcal{A}$'s queries is as follows.

Create User

When a given query ($\mathsf{ID}, psk_{\mathsf{ID}}$) is the i-th Create User query, $\mathcal{B}$ stores (ID, psk_{ID}, pk^*, $\perp$) into `List` and returns pk^* as upk_{ID}. When it is not the i-th, $\mathcal{B}$ runs (upk_{ID}, usk_{ID}) ← UserKeyGen($params$, ID). Then $\mathcal{B}$ stores (ID, psk_{ID}, upk_{ID}, usk_{ID}) into `List` and returns upk_{ID}.

Reveal Secret Key

When ID is in `List`, if $upk_{\mathsf{ID}} = pk^*$, $\mathcal{B}$ stops and outputs random bit b' and otherwise returns usk_{ID}. When ID is not in `List`, $\mathcal{B}$ returns $\perp$.

Replace

When ID is in `List`, if $upk_{\mathsf{ID}} = pk^*$, $\mathcal{B}$ stops and outputs random bit b'. If usk' = $\perp$, $\mathcal{B}$ sets usk' = usk_{ID}.Then it replaces (ID, psk_{ID}, upk_{ID}, usk_{ID}) with (ID, psk_{ID}, upk', usk').

Decrypt

When ID is in `List`, if $upk_{\mathsf{ID}} \neq pk^*$, $\mathcal{B}$ runs d_T ← Release(tpk, tsk, T) and m ← Decrypt($params$, tpk, psk_{ID}, usk_{ID}, d_T, c) and returns m. If upk_{ID} = pk^*, $\mathcal{B}$ responds as follows. If Verify($vk, c_1||c_2||c_3||\mathsf{ID}||T, \sigma$) = `reject`, then $\mathcal{B}$ returns $\perp$ to $\mathcal{A}$. Otherwise $\mathcal{B}$ runs $s_1||vk' \leftarrow$ IBE.Dec($params$, d_{ID}, c_1) $s_2||vk'' \leftarrow$ IBE′.Dec(tpk, d_T, c_2,) and issues decryption query c_3 to $\mathcal{C}$ and obtains $s_3||vk'''$. $\mathcal{B}$ returns $m = s_1 \oplus s_2 \oplus s_3$ to $\mathcal{A}$ if $vk = vk' = vk''$, and otherwise $\mathcal{B}$ returns $\perp$ to $\mathcal{A}$. When ID is not in `List`, $\mathcal{B}$ returns $\perp$ to $\mathcal{A}$.

Challenge

If $\mathcal{A}$ outputs (m_0, m_1, ID^*, T^*) as challenge, $\mathcal{B}$ runs (sk^*, vk^*) ← SigGen(λ) and randomly chooses $s_1 \in \{0,1\}^{|m|}$ and $s_2 \in \{0,1\}^{|m|}$ and runs c_1^* = IBE.Enc($params, \mathsf{ID}^*, s_1||vk^*$) and c_2^* = IBE′.Enc($tpk, T^*, s_2||vk^*$) . Then $\mathcal{B}$ computes $M_0 = [(s_1 \oplus s_2 \oplus m_0)||vk^*]$ and $M_1 = [(s_1 \oplus s_2 \oplus m_1)||vk^*]$, and issues (M_0, M_1) as $\mathcal{B}$'s challenge to $\mathcal{C}$ and obtains a ciphertext c_3^*. $\mathcal{B}$ runs σ^* = Sign($sk^*, c_1^*||c_2^*||c_3^*||\mathsf{ID}^*||T^*$) and returns $c^* = (c_1^*, c_2^*, c_3^*, \mathsf{ID}^*, T^*, vk^*, \sigma^*)$ as challenge ciphertext to $\mathcal{A}$.

Phase2

$\mathcal{B}$ responds to Create User query, Reveal Secret Key query and Replace query in the same way as in **Phase1**. $\mathcal{B}$ responds to Decrypt query as follows.The followings are done in a sequential way.

Step1

If Verify($vk, c_1||c_2||c_3||\mathsf{ID}||T||, \sigma$) = `reject`, then $\mathcal{B}$ returns $\perp$ and skips **step2∼4**.

Step2

If $vk = vk^*$, then $\mathcal{B}$ stops and outputs random bit b'.

Step3

If $c_3 = c_3^*$, then $\mathcal{B}$ returns $\perp$.

Step4

$\mathcal{B}$ responds in the same way as in **Phase1**.

Guess If $\mathcal{A}$ outputs a bit, then $\mathcal{B}$ outputs the same bit as its guess.

We consider the $\mathcal{B}$'s simulation of the response to decryption queries in **Phase2**. In the case of Verify = `reject` in **Step1**, $\mathcal{B}$ returns $\perp$ in the same way as in our Decrypt algorithm, and then it perfectly simulates the challenger in Mal.KGC game. In the case of $vk = vk^*$ in **Step2**, the event `Forge` occurs. In the case of $c_3 = c_3^*$ in **Step3**, since c_3 equals to c_3^*, the decryption of c_3 is $M_0 = [(s_1 \oplus s_2 \oplus m_0)||vk^*]$ or $M_1 = [(s_1 \oplus s_2 \oplus m_1)||vk^*]$. However, since $vk \neq vk^*$, the decryption of c is $\perp$, and then $\mathcal{B}$ simulates perfectly. In the case of $c_3 \neq c_3^*$, $\mathcal{B}$ can issue the valid decryption query c_3 to $\mathcal{C}$.

If the event `Forge` does not occurs, $\mathcal{B}$ perfectly simulates the challengers in the IND-ID-CCA$_{\mathsf{TS}}$ game. Let $\texttt{Succ}^{\mathcal{PKE}}$ denote the event that $\mathcal{B}$ wins the IND-CCA game.

We see that

$$\begin{aligned} Adv_{\Delta,\mathcal{B}}^{\mathsf{IND\text{-}CCA}} &= |\Pr[\texttt{Succ}^{\mathcal{PKE}}] - \frac{1}{2}| \\ &\geq |\frac{1}{2} \cdot (1 - \frac{1}{q}) \\ &\quad + (\Pr[\texttt{Succ} \wedge \overline{\texttt{Forge}} + \frac{1}{2}\Pr[\texttt{Forge}]) \cdot \frac{1}{q} - \frac{1}{2}| \end{aligned}$$

$$= \quad |\Pr[\mathtt{Succ} \wedge \overline{\mathtt{Forge}}] + \frac{1}{2}\Pr[\mathtt{Forge}] - \frac{1}{2}| \cdot \frac{1}{q}.$$

$Adv^{\mathsf{IND\text{-}CCA}}_{\Delta,\mathcal{B}}$ is negligible since we assume Δ is IND-CCA secure. Therefore, $|\Pr[\mathtt{Succ} \wedge \overline{\mathtt{Forge}}] + \frac{1}{2}\Pr[\mathtt{Forge}] - \frac{1}{2}|$ is also negligible. □

We see that

$$\begin{aligned}
Adv^{\mathsf{Mal.KGC}}_{\Gamma,\mathcal{A}} &= |\Pr[\mathtt{Succ}] - \frac{1}{2}| \\
&= |\Pr[\mathtt{Succ} \wedge \mathtt{Forge}] - \frac{1}{2}\Pr[\mathtt{Forge}] \\
&\quad + \frac{1}{2}\Pr[\mathtt{Forge}] + \Pr[\mathtt{Succ} \wedge \overline{\mathtt{Forge}}] - \frac{1}{2}| \\
&\leq |\Pr[\mathtt{Succ} \wedge \mathtt{Forge}] - \frac{1}{2}\Pr[\mathtt{Forge}]| \\
&\quad + |\Pr[\mathtt{Succ} \wedge \overline{\mathtt{Forge}}] + \frac{1}{2}\Pr[\mathtt{Forge}] - \frac{1}{2}| \\
&\leq \frac{1}{2}\Pr[\mathtt{Forge}] \\
&\quad + |\Pr[\mathtt{Succ} \wedge \overline{\mathtt{Forge}}] + \frac{1}{2}\Pr[\mathtt{Forge}] - \frac{1}{2}|.
\end{aligned}$$

$Adv^{Mal.KGC}_{\Gamma,\mathcal{A}}$ is negligible from **Claim 1** and **Claim 2**. This completes the proof of **Theorem 1**. □

2) Mal.Receiver *security:*

Theorem 2: If Π' is an IND-ID-CCA secure identity-based encryption scheme and Σ is an OT-sEUF-CMA secure one-time signature scheme, then Γ is a Mal.Receiver secure timed-release certificateless encryption scheme.

The proof is almost the same as that of IND-ID-CCA$_{\mathsf{CR}}$ security in [3].

3) Outsider *security:*

Theorem 3: If Π is an IND-ID-CCA secure identity-based encryption scheme and Σ is an OT-sEUF-CMA secure one-time signature scheme, then Γ is a Outsider secure timed-release certificateless encryption scheme.

The proof is almost the same as that of IND-ID-CCA$_{\mathsf{CR}}$ security in [3].

VIII. Conclusion

In this paper, we introduced a notion of TRCLE and defined Mal.KGC security, Mal.Receiver security and Outsider security. Moreover, we showed a generic construction of TRCLE in which a constructed scheme achieves those security if the primitive PKE scheme is IND-CCA secure, the primitive IBE schemes are IND-ID-CCA secure and the primitive one-time signature scheme is OT-sEUF-CMA secure.

Acknowledgment

This work was supported by JSPS KAKENHI Grant Number 26330160.

References

[1] T. May, "Timed-release crypto," Manuscript, February 1993.

[2] A. C.-F. Chan and I. F. Blake, "Scalable, server-passive, user-anonymous timed release cryptography," in *ICDCS 2005*. IEEE Computer Society, 2005, pp. 504–513.

[3] T. Oshikiri and T. Saito, "Timed-release identity-based encryption," *IPSJ Journal*, vol. 55, no. 9, pp. 1964–1970, sep 2014.

[4] ——, "Timed-release hierarchical identity-based encryption," *International Journal of Advanced Computer Science and Applications(IJACSA)*, vol. 5, no. 11, pp. 148–154, 2014.

[5] Y. Dodis and J. Katz, "Chosen-ciphertext security of multiple encryption," in *TCC 2005*, ser. Lecture Notes in Computer Science, J. Kilian, Ed., vol. 3378. Springer-Verlag, 2005, pp. 188–209.

[6] S. S. Al-Riyami and K. G. Paterson, "Certificateless public key cryptography," in *ASIACRYPT 2003*, ser. Lecture Notes in Computer Science, C. S. Laih, Ed., vol. 2894. Springer-Verlag, 2003, pp. 452–473.

[7] M. Naor and M. Yung, "Public-key cryptosystems provably secure against chosen ciphertext attacks," in *STOC '90*. ACM, 1990, pp. 427–437.

[8] A. Shamir, "Identity-based cryptosystems and signature schemes," in *CRYPTO '84*, ser. Lecture Notes in Computer Science, G. R. Blakley and D. Chaum, Eds., vol. 196. Springer-Verlag, 1985, pp. 47–53.

[9] D. Boneh and M. K. Franklin, "Identity-based encryption from the Weil pairing," *SIAM Journal on Computing*, vol. 32, no. 3, pp. 584–615, 2003, a preliminary version appeared in *CRYPTO 2001*, 2001.

[10] R. C. Merkle, "A digital signature based on a conventional encryption function," in *A Conference on the Theory and Applications of Cryptographic Techniques on Advances in Cryptology*, ser. CRYPTO '87. London, UK, UK: Springer-Verlag, 1988, pp. 369–378. [Online]. Available: http://dl.acm.org/citation.cfm?id=646752.704751

Vehicle Embedded Data Stream Processing Platform for Android Devices

Shingo Akiyama*, Yukikazu Nakamoto*, Akihiro Yamaguchi†, Kenya Sato‡, Hiroaki Takada§
*Graduate School of Applied Informatics, University of Hyogo, Japan
†Center for Embedded Computing System, Nagoya University, Japan
‡Mobility Research Center, Doshisha University, Japan
§Graduate School of Information Science, Nagoya University, Japan

***Abstract*—Automotive information services utilizing vehicle data are rapidly expanding. However, there is currently no data centric software architecture that takes into account the scale and complexity of data involving numerous sensors. To address this issue, the authors have developed an in-vehicle data-stream management system for automotive embedded systems (eDSMS) as data centric software architecture. Providing the data stream functionalities to drivers and passengers are highly beneficial. This paper describes a vehicle embedded data stream processing platform for Android devices. The platform enables flexible query processing with a dataflow query language and extensible operator functions in the query language on the platform. The platform employs architecture independent of data stream schema in in-vehicle eDSMS to facilitate smoother Android application program development. This paper presents specifications and design of the query language and APIs of the platform, evaluate it, and discuss the results.**

Keywords*—Android, automotive, data stream management system

I. Introduction

Automotive information services utilizing vehicle data are rapidly expanding. Several standardizations have been put into place for the rapid deployment of such services. Vehicle data interfaces such as OpenXC and Mirrorlink have recently been standardized and are expected to become more popular, despite the fact that existing built-in car navigation systems use proprietary and closed vehicle data. OpenXC defines APIs that provide diagnostic data from a controller area network (CAN) bus in a vehicle network [1]. The Car Connectivity Consortium has recently standardized Mirrorlink, which defines interfaces that connect smartphones with vehicle information [2]. Google and Apple announced software platforms for automotive information services, Android Auto[3] and CarPlay[4], respectively. Android Auto provides functionalities with which Android devices communicate to a vehicle. CarPlay is an iOS virtual machine on top of OS in an in-vehicle system and enables communication with iOS devices.

Automotive control is now undergoing the same technology trends described above. Intelligent control systems have become popular due to their sophisticated, safe, and environmentally friendly control exploiting numerous types of data from a vehicle itself, its surroundings, and other vehicles. Typical systems that employ such technologies include pre-crash safety systems, adaptive cruise control, lane departure warning systems, and intelligent parking assist systems. These systems collect environmental data from sensors in a vehicle and make decisions on the basis of this data to control the vehicle on behalf of the driver. Jones used information obtained from multiple on-board sensors to perform evasive steering and, when collision is unavoidable, to activate brake intervention to dampen the impact, thus decreasing damage [5]. Such intelligent control systems acquire data from many sensors, such as cameras and millimeter-wave radar. Google and Urban Challenge, which is a competition funded by the Defense Advanced Research Projects Agency, revealed that self-driving in urban areas is both feasible and safe in terms of autonomous driving [6], [7].

There are also more advanced information and control techniques for automotive systems and services that use data through communication. The automotive industry has been studying cooperative intelligent transport systems (C-ITSs) to improve transport safety, productivity, and reliability by using data collected through vehicle-to-infrastructure [1] (V2I) and vehicle-to-vehicle (V2V) communications, as well as GPS from outside the vehicle and data collected on-board [8][9][10]. Thus, C-ITS technologies enhance automotive information and control systems.

AUTOSAR [2], which is an automotive software standardized organization, recently proposed a component-based software platform and software development methodologies to address the growing scale and complexity of automotive control software [11]. AUTOSAR does not, however, take into account the scale and complexity of data involving numerous sensors. Moreover, application integration is becoming more complex because access to one sensor requires communication with an application that manages the sensor when each application manages sensors by itself. When information from multiple sensors is integrated or when new sensors or algorithms need to be added, the software architecture needs to be redesigned and reorganized. Those technologies mentioned above do not provide solutions for those problems.

In response to these issues, the authors have researched and developed a data centric software architecture for automotive systems. The evaluation results of both a database management system (DBMS) and a data stream management system (DSMS) showed advantage of the latter system because the DSMS can already efficiently handle continuous incoming data such as vehicle sensor data [12]. The above observation

[1]Infrastructure is a specific term in ITS that refers to the roads, centers, and facilities around vehicles.

[2]http://www.autosar.org/

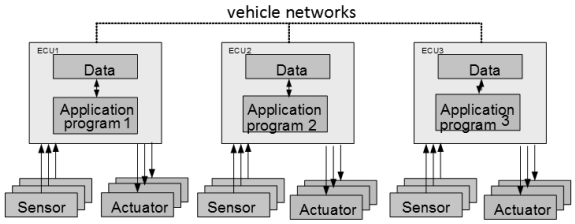


Fig. 1: Current architecture in an automotive system.

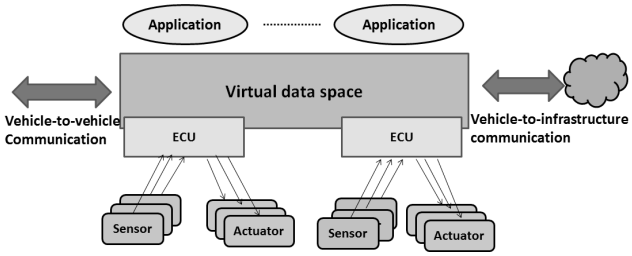


Fig. 2: Data centric architecture in an automotive system

becomes a basis for developing a data-stream management system for an automotive embedded system (eDSMS)[13], [14]. In addition, a data-stream-based local dynamic map (LDM) was previously proposed [15]. The data-stream-based LDM consists of layered data, including geography, circumjacent vehicle status road conditions, and congestion and weather circumstances and is a key technology in C-ITS.

Providing the data stream functionalities to drivers and passengers are highly beneficial. This paper describes a vehicle embedded data stream processing platform for Android devices. The rest of this paper is organized as follows. A brief overview of the data centric architecture and data stream management system for embedded systems (eDSMS) is presented in Sec. II. Sections III and IV describe the architecture and design, respectively, of the vehicle embedded data stream processing platform for Android devices whose target is vehicle information systems. SectionV shows a demonstration program utilizing the platform. Evaluation results are presented in Sec.VI. Related works are briefly discussed in Sec. VII and conclusions are described in Sec. VIII.

II. Data Stream Management System for Vehicle Embedded System: eDSMS

This section describes the data centric software architecture based on data-stream processing for automotive systems. The development of these systems has been discussed in previous works [12], [13], [14].

A. Background

The architecture of current automotive systems is shown in Fig. 1. Programs in an electrical control unit (ECU) [3] obtain data from numerous sensors, process the data, and output commands to actuators to provide control for automotive systems, including control and information systems. Sensor data are duplicated and processed in application programs in multiple ECUs because each application program has to process sensor data in its own ECU. The cost of developing and integrating application programs increases dramatically if the number of sensor data types increases. This occurs because many sensors are fixed to a vehicle or numerous data come from outside the ego-vehicle [4].

Figure 2 shows a proposed data centric software architecture for an automotive system. This architecture provides virtual data space for data not only in a vehicle but also from other vehicles and infrastructures, thus hiding the data's origin. This architecture separates sensors from application programs and provides common access methods for sensor data, which is managed collectively in the logical data space. Moreover, the proposed architecture increases opportunities for sensor fusion, which enhances one piece of sensor data with other sensor data.

To determine the feasibility of the data-centric software architecture in vehicle software, they evaluated a DBMS and a DSMS. The feasibility study used two application programs: adaptive cruise control and intelligent parking assist. The results of the evaluation demonstrated that DBMS is superior at processing queries featuring large amounts of data at low frequency while DSMS is superior at processing queries featuring small amounts of data at high frequency. This results in adopting DSMS in the data centric architecture of in-vehicle systems because most data in automotive systems are continuous sensor-generated data and have short lifetimes. At the same time, the architecture use DBMS for static data such as map information and convert the static data into stream data to pass the data on to DSMS.

B. In-vehicle eDSMS

An embedded DSMS (eDSMS) is suitable for embedded systems, especially in-vehicle embedded systems. Note that software in embedded systems must be particularly customized because CPU and memory capacities are limited and so real-time processing and high reliability of the systems are required.

This section presents an overview of DSMS. The DSMS input is stream data. DSMS obtains data in the stream specified by a query and then outputs that data as a stream. The DSMS query is issued for the data stream and is executed continuously, unlike a query in DBMS. There are two types of query language in DSMS: SQL-like declarative languages and procedural languages, specifically, dataflow languages [16][pp.723-743]. In the SQL-like query languages, a user specifies selection predicates from streamed input data. Stream data are then converted into relations in sliding windows in DSMS. A query is executed over the sliding window similarly to a conventional SQL query. The result of the query is then converted into a stream again. In the dataflow query languages, a user specifies a query with a dataflow graph, where a node is a query operator (discussed later) and an edge is a stream. Data from the input stream flows in the dataflow query. An operator processes data in the dataflow query and detects the specified data. In this way, a user can describe a query procedure explicitly in the dataflow query languages.

[3] An electric control unit (ECU) is a computer used for vehicle control.

[4] An ego-vehicle means the vehicle that is being focused upon.

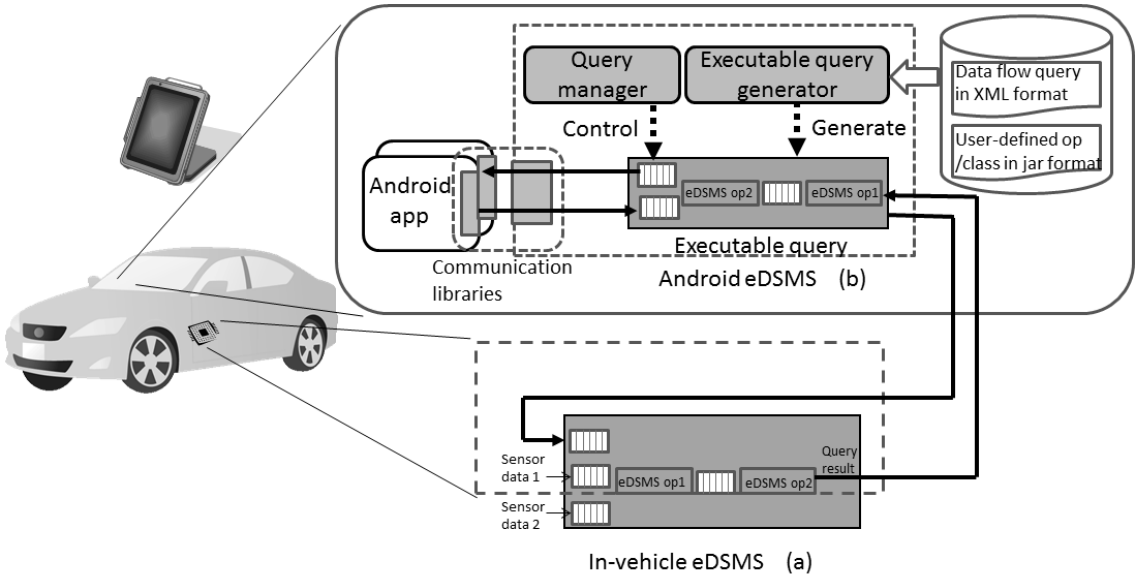


Fig. 3: Android eDSMS and in-vehicle eDSMS.

The in-vehicle eDSMS has three features of embedded systems: dataflow query language, static query processing, and optimization. First, a query language in the in-vehicle eDSMS is a dataflow language that is similar to the one in Borealis[17]. This is because a query in the dataflow language is more flexible in terms of customizing and tuning the processing for automotive applications. Second, a dataflow language constructs a hierarchical structure of data easily from physical layers to more abstract layers. Finally, a user can reuse data in the arbitrary level of the dataflow query, which avoids redundant usage of data.

In the automotive field, in-vehicle application programs do not change dynamically. This means there is no updating, adding, or deleting the application programs because the programs are fixed to guarantee reliability and safety after the long-term verification and validation of the programs. Thus, the query processing in in-vehicle eDSMS does not change dynamically either. The query, whose actual representation is XML-form, is converted into C/C++ source programs that are compiled into run-time routines on a PC, embedded into the ECU in a vehicle, and executed as part of the vehicle's programs. The execution time of a query is predictable in in-vehicle eDSMS. This predictability property is very important in real-time systems such as vehicle systems because they need to be guaranteed to finish their processes.

There are several optimization techniques in in-vehicle eDSMS query processing for reducing the processing overhead and ROM/RAM usage, including deleting operator dynamic linking, linking selected operator modules only, and decreasing the number of tasks used in in-vehicle eDSMS runtime.

III. eDSMS in Android platform

A. Requirements

Providing the data stream functionalities to drivers and passengers are highly beneficial. The role of an embedded data dream management system for an Android platform in a device (Android eDSMS or AeDSMS) is to provide not only straightforward usage of the in-vehicle data stream to drivers and passengers but also various usage data streams between in-vehicle and Android devices. The requirements of Android eDSMS are considered in the following cases:

- Case 1: Presenting services to drivers and passengers utilizing the data stream in the in-vehicle eDSMS. A simple application is to inform a driver of warnings and cautions. Another application is to retrieve information suitable for the driving situation (e.g., location, time, weather) from the stream data from the in-vehicle eDSMS and present it to the driver.
- Case 2: Sending input information from Android applications as data stream or parameters to an AeDSMS operator in the dataflow query to in-vehicle eDSMS. For example, a driver may control a vehicle or inform other vehicles of a traffic condition in the form of sensor data to the in-vehicle eDSMS.
- Case 3: Executing part of the eDSMS dataflow query from the in-vehicle in the Android platform. The purpose of this execution is to debug a query of the in-vehicle EDSMS in a more convenient programming environment or to offload query processing in an Android device to better utilize resources within the device itself. In offload usage, although the processing time of the query becomes shorter, predictability is lost because the operating system of the offloaded Android device has a general purpose OS.

B. Prototype

A prototype of eDSMS for Android devices has the following features for the feasibility evaluation [18]. The target

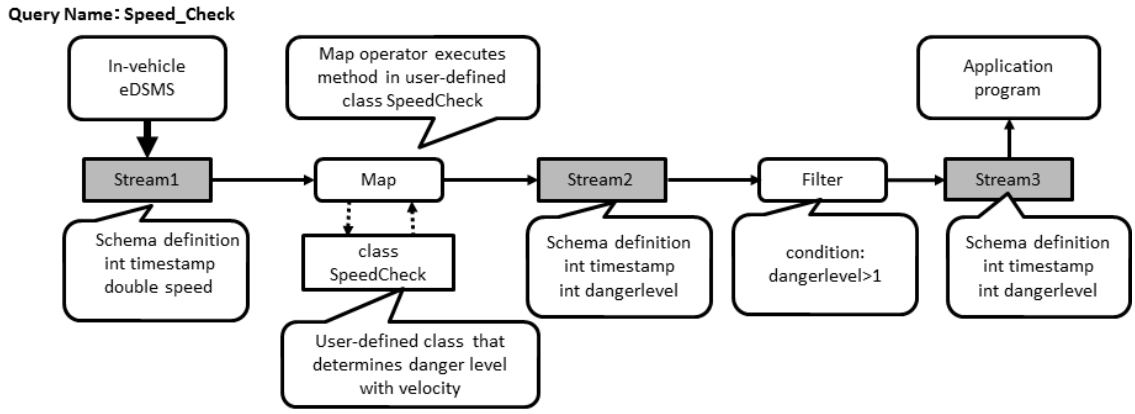


Fig. 4: Query example.

of the prototype is Case 1 from the previous subsection. In the Android eDSMS prototype, data is filtered from the in-vehicle eDSMS data stream and passed on to the Android eDSMS. The filtering is specified by an SQL-like query. The development process of an Android application using the prototype is as follows.

1) A developer creates a schema file in an XML format that defines the field information of the stream data acquired from the in-vehicle eDSMS.
2) The AeDSMS builder in a PC obtains a schema file as an input and generates the Java source codes of a query in the prototype, including the class of data received from the in-vehicle eDSMS.
3) A developer integrates application source codes with the query source codes and compiles them to produce an application program in the PC for an Android device.

If the schema of a data stream from the in-vehicle eDSMS changes, it is necessary to generate the query source codes and compile source codes again. This is neither convenient nor flexible.

C. Architecture of Android eDSMS

On the basis of the experience with the prototype development, an enhanced embedded data stream management system for an Android device has the following three features. The purpose is to address the issues mentioned above in Cases 1 and 2 and to add features to improve the developing efficiency from the perspective of the developer.

- Flexible query processing with a dataflow query language and extensible function of operators in the query language
- Facilitation of Android application program development with architecture independent of data stream schema in in-vehicle eDSMS and class libraries to hide implementation details
- Capability of various usages in multiple Android applications by providing query processing as an Android service

Android DSMS is part of a two-layer structure for an embedded data stream platform for automotive systems, shown in Fig. 3. Figure 3 (a), the bottom, is an in-vehicle eDSMS in ECUs, and Fig. 3 (b), the top, is an Android eDSMS. In this structure, the in-vehicle eDSMS obtains sensor data, processes the data in the form of general purpose usage, and sends it to the Android platform. The Android eDSMS receives the data stream from the in-vehicle eDSMS and provides various usages of the data stream between in-vehicle and Android devices, such as driving situation services including location, time, and weather.

D. A dataflow query of Android eDSMS

AeDSMS adopted a dataflow query language in the data stream query the same as the one in in-vehicle eDSMS, which is similar to the one in Borealis [17]. There are two types of dataflow query files in the XML format: a schema file and a query file. The schema file contains schema information of the stream in a query and the query file describes operators and connections between operators used in the query. Built-in query operators in AeDSMS are listed in Table I, which are also the same as in-vehicle eDSMS. An operator has input streams, output streams, and parameters required for execution. AeDSMS reads both files, translates them into an executable query, and executes the query in an Android device (see Fig. 3). A developer can extend the functions of an operator, which means these functionalities provide developers with easier, more flexible application development.

An example of a query is shown in Fig. 4. This is a simple query to determine the danger level according to the vehicle speed and to select only those levels higher than 1. Below is an excerpt of the query in the XML format.

TABLE I: Operators in Android eDSMS.

Operator	Description
Filter	read data from an input stream, perform filtering in accordance with the condition data, and write the result to an output stream
Map	read data from an input stream, perform the specified method on it, and write it to an output stream
Unite	read data from multiple input streams, merge and write it to an output stream
Join	read data from an input stream, hold it for certain period of time, combine the data held by the specific conditions, and write it to an output stream
Aggregate	read data from an input stream, hold it for certain period of time, perform aggregate functions on the data, and write it to an output stream

```
<query name="speed_check">--(1)
<box name="test_map" type="Map">--(2)
  <in stream="Stream1" />
  <out stream="Stream2" />
  <parameter name="speed_check"
    class="SpeedCheck"/>--(3)
</box>
<box name="speed_filter" type="Filter">
  <in stream="Stream2" />
  <out stream="Stream3">
  <parameter name="expression_0"
    value="dangerlevel>1" />--(4)
</box>
</query>
```

A `query` element represents a query. A query name is defined in a `name` attribute (1). In a `box` element, an operator is defined and a `type` attribute specifies an operator type (2). This example uses `Map` and `Filter` operators. The `box` element consists of an operator name, input and output streams for the operator, and operator parameters. A class `SpeedCheck` is specified as a user defined class called by the Map operator (3). The Filter operator specifies a condition where data is passed to Stream 3 only if the value of the variable `dangerlevel` representing the danger level exceeds 1 (4).

AeDSMS provides two operator extensions: an extensible operator with a user defined class and a user defined operator. A developer can write a class where a method is called from an operator to extend and change the process and execution conditions of that operator. Moreover, a developer can define and add a new operator for implementing a flexible query.

Three methods exist in AeDSMS to communicate with in-vehicle eDSMS: TCP, UDP, and Bluetooth. A developer can select the method by specifying it in a query file.

IV. Design of Android eDSMS

The Android eDSMS has the following parts (see Fig.3):

Executable query generator:
: The executable query generator reads the XML format queries, generates executable queries, and classes objects that have field information of the data in the data stream to receive the data in an Android device, as shown in Fig. 3(b).

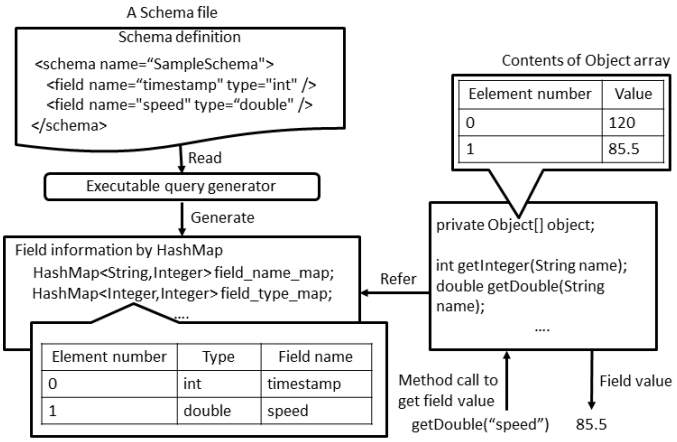


Fig. 5: Stream data access using HashMap generated from query file.

Executable query:
: The executable query is the run time part of the query.

Communication libraries:
: Communication libraries hide the detail of the inter-process communication and structure of AeDSMS.

Query manager:
: The query manager manages the executable query.

From the above architecture, especially the executable query generator, a developer can change the schema in an input data, the process, and the communication method simply by rewriting the query file without generating any AeDSMS code in a PC. A query in AeDSMS can dynamically change in order to provide new services while the program of in-vehicle eDSMS is embedded in a static structure. Also, it is possible for multiple Android applications to use AeDSMS through AeDSMS's Android service.

A. Executable Query Generator

The executable query generator reads the schema in a schema file and the query in a query file and then generates the hash map at run time, as shown in Fig. 5, instead of producing class files corresponding to the data from the in-vehicle eDSMS during the development, as described in Sec. III-B. At this time, a class `AeDSMSFieddInfo` is generated that contains the field name, type and element numbers. Those field are related in the hash map with `HashMap`. At this time, the field name, type, and element numbers are related in the hash map with `HashMap`. Data received from in-vehicle eDSMS is initially stored in an `Object` type array. A class `AeDSMSTupleData` corresponding to the tuple data received from the in-vehicle eDSMS has `getDouble()` and `getInteger()` methods that cast the data `Object` type to a type specified by the method name and retrieve the specified data with the hash map. Moreover, A class `AeDSMSTupleData` provides methods `putDouble()` and `putInteger()` to put data into a tuple data. Therefore, no modification of the Android program is required, even if the data schema obtained from the in-vehicle eDSMS has changed.

B. Extension of an operator

AeDSMS provides two operator extensions: an extensible operator with additional class and a user defined operator. For the extension of an operator, a developer can create a class that implements the interface `AeDSMSUserDefineClassBase`, which has the following methods.

```
interface AeDSMSUserDefineClassBase {
  abstract public AeDSMSTupleData[]
    execution(AeDSMSTupleData[] t,
      Object[] args);
  abstract public String getName();
}
```

An abstract method `execution()` is implemented in the operator with parameters `AeDSMSTupleData[] t` and `Object[] args`. The first parameter `t` is data for executing the extension and the second one is additional.

Another method for an operator extension is a user defined operator. Here, a developer writes a user defined operator in a q query file. In the operator definition, a `type` attribute is the name of the class where the user defined class is written.

```
<box name="test_map" type="UserOp">
  <in stream="Streamx" />
  <out stream="Streamy" />
  <parameter "UserOp parmeters"/>
</box>
```

A user defined operator can be defined with inheritance of an abstract class `AeDSMSOperatorBase`. The class `AeDSMSOperatorBase` has the following methods. In defining the user defined operator, a developer must write how to parse the parameters of the operator defined in the dataflow query file as above. To do that, the developer writes `parse()` in the class definition of a user defined operator. A method `parse()` is called when loading the dataflow query file containing a user defined operator with a `loadXMLQuery()` execution. A method `execution()` is called from the executable query when the operator is executed. A method `isExecutable()` returns information on whether the operator is executable.

```
abstract class AeDSMSOperatorBase {
  public abstract void
    parse(Nodelist node_list);
  public abstract void execution();
  public abstract boolean isExecutable();
}
```

A developer writes a class for extension of an operator and a user defined operator, compiles it, and stores it in Jar file format in an Android device. The executable query generator extracts class information from a user defined class in the Jar format and translates it into an executable Dalvik dex (Dalvik EXecutable) format, which is available on Android. A `dx` tool in the Android SDK performs the translation. We assume that classes for frequently used general processing can be prepared in advance.

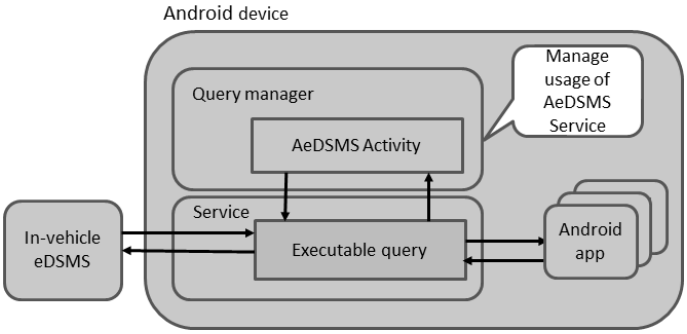


Fig. 6: Multiple process access to Android eDSMS by service.

The operators listed in Table I can also implement the `execution` method. A scheduler in the executable query calls the `execution()` method of each class instance of the operator when the method is executable.

C. Query management

An Android application can read multiple queries and make selections from within those queries. AeDSMS prepared a class `QueryManager` to manage the multiple queries. `QueryManager` also manages the multiple threads needed for a query execution. An excerpt of the member methods is shown below. `QueryManger` holds a query in `ArrayList`. A method `addQuery ()` adds a query in the List. A method `startQuery()` starts execution of the specified query.

```
class QueryManager{
  public void addQuery(AeDSMSQuery query);
  public void deleteQuery(String name);
  public void startQuery(String name);
  public void cancelQuery(String name);
  public AeDSMSQuery getQuery(String name);
}
```

D. Usage from multiple processes by service.

If activity in an Android application creates an AeDSMS instance and runs it, it means that the AeDSMS instance is equal to the number of applications that exist. Moreover, utilizing a single AeDSMS from multiple applications or remote usage of other Android devices is desired. Therefore, AeDSMS is separated from an application to solve this problem and is executed as a separate process with the Android service, shown in Fig. 6. A developer can either select an Android service in an application or perform AeDSMS in an application, where AeDSMS is executed as part of an activity.

E. Process communication and communication library

When running the Android service and an application in separate processes, using inter-process communication for exchanging data is necessary. Android eDSMS implements process communication with Messenger/Handler, which is one of inter-process communication mechanism in Android.

AeDSMS provide a communication library to hide the details of the inter-process communication and programming

in an Android application, rendering it independent of the inter-process communication. One class of the library is shown below as an example.

```
class AeDSMSComm{
  public void loadXMLSchema(String name);
  public void loadXMLQuery(String name);
  public void startAeDSMSQuery(String name);
  public void cancelAeDSMSQuery(String name);
  public AeDSMStream
    getAeDSMStream(String name);
}
class AeDSMSStream {
  public AeDSMSTupleData
    getAeDSMSTupleData();
  public void
    putTupleData(AeDSMSTupleData t);
}
```

A class `AeDSMSComm` provides the following methods.

- A method `loadXMLSchema` loads a `name` schema file and a method `loadXMLQuery` loads a `name` query file in the XML format into AeDSMS and generates the executable query.
- A method `startQuery` starts execution of the `name` query in the loaded query file and `cancelQuery` stops the execution.
- A method `getAeDSMStream` returns the result of the query in the named output stream in the query file.

A class `AeDSMSStream` contains a result of the query that is specified in the named output stream in the query file. The Android application program calls a method `getInteger()` or `getDouble()` with the field name a developer wants to get from the tuple and obtains the value of the tuple as

```
AeDSMSComm ac = new AeDSMSComm();
ac.loadXMLSchema("speed_check_schema.xml");
ac.loadXMLQuery("speed_check.xml");
ac.startAeDSMSQuery("speed_check");
              :
AeDSMSStream s = getAeDSMSStream("Stream3");
AeDSMSTupleData t = s.getAeDSMSTupleData();
int ts = t.getIneger("timestamp");
int dl = t.getInteger("dangerlevel");
```

When a developer send a tuple data to an input of an AeDSMS executable query, he or she generates an input stream for AeDSMS with `getAeDSMSStream` and a tuple data of `AeDSMSTupleData`. A method `putDouble()` puts a value to the specified field and `addTupleData()` adds the tuple data to the stream. Suppose that AeDSMS sends speed control values to the in-vehicle eDSMS. A speed value 50.0 is set to "speed" field in a tuple and the tuple is output to the `Stream4`, which is connected to the in-vehicle eDSMS.

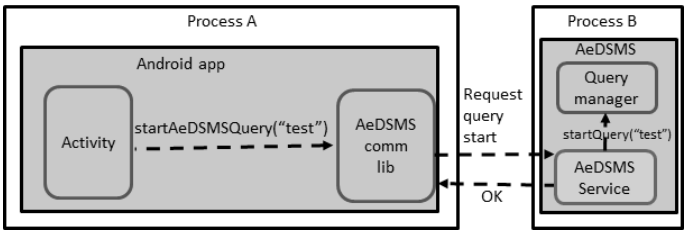


Fig. 7: Hiding implementation details by communication library.

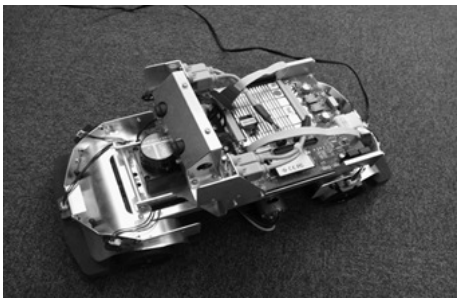


Fig. 8: Robocar 1/10.

```
AeDSMSStream s = getAeDSMSStream("Stream4");
AeDSMSTupleData t =
  new AeDSMSTupleData(finfo);
t.putDouble("speed", 50.0);
s.addTupleData(t);
```

A developer can write an application using `StartAeDSMSQuery()` and `getAeDSMSStream()`, as shown in Fig. 7, without being aware of the process communication.

V. Developing Android application using AeDSMS

This section presents a demonstration application using AeDSMS in NEXUS 7 with ASUS and Google. The vehicle in this demonstration is a miniature of ZMP's RoboCar 1/10[5], shown in Fig. 8. The Android application has a sensor information display function, a battery power display function, and a speed-meter display function.

Figures 9 and 10 show screen shots of the application. Figure 9 provides basic information of the driving. Steering angle and infrared sensor information are displayed using animation on the left side of the screen. A user can tap the speed on the screen of Fig. 9 and transit to the screen shown in Fig. 10, where the current speed measured by a meter is displayed. A developer can thus write various kinds of Android applications more easily and productively.

VI. Evaluation

This section describes performance evaluation of the Android eDSMS on NEXUS7 with the program in Fig. 4. The

[5]http://www.zmp.co.jp/?lang=en

TABLE II: Specification of NEXUS 7

CPU	NVIDIA Tegra 3 (1.3 GHz)
Memory	1GB
OS	Android 4.4.2

Fig. 9: User interface of Android application using Android eDSMS.

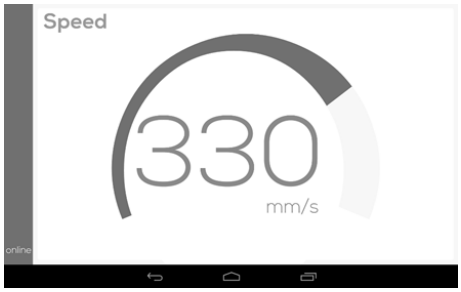


Fig. 10: Another user interface of Android application using Android eDSMS.

TABLE III: Measurement results

processing time with stream operators using AeDMSM	150 μs
processing time with empty operators using AeDMSM	90 μs

measurement time was the end-to-end duration with the operators from time at `Stream 1` to time at `Stream 3` in Fig. 4 as well as the duration of empty operators, which indicates the overhead of the query execution time in AeDSMS. This measurement was performed 100 times and the averages was calculated. The results are shown in TABLE III. These values show that the overhead of the query execution time is relatively small compared with the execution time of the operator, and thus demonstrating the feasibility of implementing AeDSMS.

The overhead of a query in a single process as an "Activity" and query execution using the inter-process communication as a "Service" are different because of the separate processes used. The execution contains an 'empty' query with one empty operator and two streams to measure this difference and found that it takes 4.70 ms to execute a method `loadXMLQuery();` in the single process and 8.76 ms using the AeDSMS service. This measurement was performed 100 times and the averages was calculated. Those executions involve garbage collection time. This difference is the overhead of the inter-process communication as a "Service." A developer should choose to run a query as a single process if there will be no multiple application usages of AeDSMS.

If an Android program consumes memory, garbage collection (GC) occurs and pauses program execution. The execution contains the same 'empty' query and observed its GC when a tuple with one `int` field was input at intervals of 10 ms over 500 s. GC occurred six times during the execution and the average pause time was 18.5 ms. GC is unavoidable in Java and Android, however, one way to prevent GC is that a program generates and reserves tuples that are used before program execution, and gets a tuple from the tuple reservation instead of the instance creation at runtime. In this case, there is a limit of the number of received tuples within a certain time.

VII. RELATED WORKS

For general-purpose DSMS, prototype and commercial systems of Aurora [19] and its successor Borealis [17] and STREAM [20] have been developed. Aurora and Borealis adopt a dataflow language as a query language while STREAM has an SQL-like query language. In the finance field, DSMSs are used in applications related to algorithmic trading and financial monitoring. In the case of algorithmic trading, it is necessary to reduce the response time of query processing, as this directly affects profit. In addition, finance-based DSMSs must update queries immediately when the algorithm is updated. These systems enhance features, such as providing several types of windows between stream operators for real world applications. However, all stream operators, queues (variable length), and TCP communication are embedded as the standard executable code, which leads to larger code size. Part of receiving a query result in an application can be executed in the same thread in the commercial version of STREAM to reduce receiving time latency. Conventional DSMSs are also often applied across the Internet. Such DSMSs process packets as a stream, requiring high throughput rather than adherence to any deadline, unlike in the automotive field. In addition, Internet-based DSMSs often use overlay networks, which are different from in-vehicle networks.

DSMS has started being utilized in embedded systems. The first utilizations have been in the automotive field. Schweppe et al. proposed on-board stream processing for engineering testing and diagnosis in vehicle systems [21]. One of their main features was the adaptation of the behavior of data stream processing in diagnosis when critical events occur, e.g., when the reading rate of sensor data increases. However, their streaming platforms cannot schedule data processing so as to meet deadlines. StreamCars, which is most similar to in-vehicle eDSMS in terms of purpose, proposed a software development platform for vehicle embedded systems. Although StreamCars provides sensor fusion operators, the performance and implementation have not been described in detail.

The Cooperative Cars (CoCar) project at Aachen University is developing a data stream mining platform for automotive systems [22]. Examples of its application include queue-end detection and traffic state estimation. These are processed on the server-side rather than in an automotive embedded system. Although such applications must perform spatial operations to determine which road a vehicle is driving on, the deadline constraint is looser than in driving assistance systems. Unlike

in-vehicle eDSMS, the CoCar platform processes spatial operations using an RDBMS. Data quality (DQ) is important in DSMS, but it is not extensive. Their group also generalized the DQ of a data stream using ontology [23].

Researching real-time scheduling of DSMS is popular because real-time processing from inputs and to outputs is a key property in embedded systems such as automotive systems. [24] presented a real-time scheduling algorithm to guarantee the required quality-of-service level in embedded DSMS. They define the quality-of-service level and resources needed for computation by DSMS operators and provide a framework where a user negotiates the quality in DSMS. Son et al. proposed a periodic query model for real-time applications and an admission control mechanism for an overload situation with irregular stream data arrival [25]. As another scheduling approach, a preemptive rate-based operator scheduling has been proposed [26]. The rate-based scheduling enables earlier execution operators on an operator path in the data stream to perform processing with higher priority. An operator with higher priority can be immediately executed by preempting the current executing operator if the operator is ready. In [27], a task processes data on an operator path in a dataflow query and an operator scheduling algorithm is examined in which a task is earlier executed corresponding to data in the stream with the earliest deadline among the waiting data.

In the second, data processing in a sensor network can be regarded as a data stream [28]. DSMSs are applied to applications such as traffic monitoring and environmental monitoring. As in the embedded field, a small footprint is required because low-specification nodes are often used. Additionally, many applications require distributed processing, and minimal network usage is necessary to preserve battery power and save precious network bandwidth. However, these networks are basically peer-to-peer, which differ from in-vehicle networks. Several previous works based on sensor network ideas, resource saving DSPSs that can be installed in embedded systems, have been developed for the purpose of aggregating and monitoring sensor data [29], [30]. Müller's DSMS[30] is for a wireless sensor node. A query is registered statically and converted into intermediate codes executed on a virtual machine. In the virtual machine, 37 instructions are borrowed from a Java virtual machine and 27 instructions are specified for the data stream processing. They adopted a declarative query language, making it possible to increase the abstraction level and enable in-network programming in a sensor network, reduce the program size in a sensor node, and easily reprogram sensor nodes.

Gigascope [31] is a DSMS for the network equipment in the base station. A query is statically registered and converted into C and C++ source codes, the same as in-vehicle in-vehicle eDSMS. Details have not been published for Gigascope, and there is no description of the optimization of in-vehicle eDSMS.

VEDAS [32] and Minefleet [33] are DSPSs that mainly target mobile computing devices. Their applications relate to data stream mining, i.e., vehicle-health monitoring and driver characterization. They distribute stream processing among mobile computing devices so as to reduce battery usage and wireless communication. However, they are not intended for in-vehicle networks.

VIII. Conclusion

This paper presented a vehicle embedded data stream processing platform for Android devices to provide the data stream functionalities to drivers and passengers with many benefits. The platform enables flexible query processing with a dataflow query language and extensible operator functions in the query language in the platform. The platform has an architecture independent of data stream schema in in-vehicle eDSMS to facilitate Android application program development. Future work includes adopting SQL-like query language for programming that is familiar to Android application programmers and asynchronous query execution for the data stream.

Acknowledgments

The authors thank Masanori Okamoto and Mohanmed Bhuiya for the prototype development of the Android eDSMS and Shinichi Ito, Naoyuki Shiba, Hideteru Shimada, Toshihiko Sugawara, and Naoya Suzuki for the in-vehicle eDSMS development.

Finally, the authors thank Yoshitaka Nakagawa, University of Hyogo, for creating GUIs for the demonstration application program using Android eDSMS.

The present study was supported in part by MIC SCOPE 121806015, JSPS KAKENHI Grant Numbers 25240007 and 24500045. This work was partly done in Education Network for Pratical Information Technologies (enPiT), Japan.

References

[1] OpenXC, "The OpenXC Platform." [Online]. Available: http://openxcplatform.com/

[2] Car Connectivity Consortium, "MirrorLink 1.0 specification," Tech. Rep., 2013.

[3] Google, "Android Auto," 2014. [Online]. Available: http://www.android.com/auto/

[4] Apple, "Apple CarPlay," 2014. [Online]. Available: http://www.apple.com/ios/carplay/

[5] W. D. Jones, "Keeping cars from crashing," *IEEE Spectr.*, vol. 38, no. 9, pp. 840–851, 2011.

[6] M. Buehler, K. Iagnemma, and S. Singh, Eds., *The DARPA Urban Challenge: Autonomous Vehicles in City Traffic*. Springer, 2010.

[7] E. Guizzo, "How Google's Self-Driving Car Works," 2011. [Online]. Available: http://spectrum.ieee.org/automaton/robotics/artificial-intelligence/how-google-self-driving-car-works

[8] G. Toulminet, J. Boussuge, and C. Laurgeau, "Comparative synthesis of the 3 main European projects dealing with Cooperative Systems (CVIS, SAFESPOT and COOPERS) and description of COOPERS Demonstration Site 4," in *International Conference on Intelligent Transportation*, 2008.

[9] ETSI, "Intelligent Transport Systems (ITS): Communications Architecture,," 2010.

[10] D. Zhang, H. Huang, M. Chen, and X. Liao, "Empirical study on taxi GPS traces for Vehicular Ad Hoc Networks," in *Proc. 2012 IEEE International Conference on Communications*, 2012, pp. 581–585.

[11] S. Fürst, J. Mössinger, S. Bunzel, T. Weber, F. Kirschke-Biller, P. Heitkämper, G. Kinkelin, P. Citroën, K. Nishikawa, and K. Lange, "AUTOSAR - A Worldwide Standard is on the Road," in *Proc. 14th International VDI Congress Electronic Systems for Vehicles*, 2009.

[12] M. Yamada, K. Sato, and H. Takada, "Implementation and evaluation of data management methods for vehicle control systems," in *Proc. IEEE 74th Vehicular Technology Conference*, 2011, pp. 1–5.

[13] S. Katsunuma, S. Honda, Y. Watanabe, Y. Nakamoto, and H. Takada, "Real-time-aware Embedded DSMS Applicable to Advanced Driver Assistance Systems," in *Proc. 33rd IEEE Symposium on Reliable Distributed Systems Workshops*, 2014, pp. 5–10.

[14] A. Yamaguchi, Y. Nakamoto, K. Sato, Y. Ishikawa, Y. Watanabe, S. Honda, and H. Takada, "AEDSMS: Automotive Embedded Data Stream Management System," in *31st IEEE International Conference on Data Engineering*, 2015 (accepted).

[15] K. Sato, H. Shimada, S. Katsunuma, A. Yamaguchi, M. Yamada, S. Honda, and H. Takada, "Stream LDM : local dynamic map (LDM) with stream processing technology," Doshisha University, Tech. Rep., 2012.

[16] M. T. Özsu and P. Valduriez, *Principles of Distributed Database Systems*, 3rd ed. Springer, 2011.

[17] D. J. Abadi, Y. Ahmad, M. Balazinska, J.-h. Hwang, W. Lindner, A. S. Maskey, A. Rasin, E. Ryvkina, N. Tatbul, Y. Xing, and S. Zdonik, "The Design of the Borealis Stream Processing Engine," in *Proc. Second Biennial Conference on Innovative Data Systems Research*, 2005, pp. 277–289.

[18] Y. Nakamoto, M. Okamoto, M. Bhuiya, A. Yamaguchi, K. Sato, S. Honda, and H. Takada, "Android Platform based on Vehicle Embedded Data Stream Processing," in *2013 IEEE 10th International Conference on Ubiquitous Intelligence & Computing and 2013 IEEE 10th International Conference on Autonomic & Trusted Computing*, 2013, pp. 48–55.

[19] D. J. Abadi, Y. Ahmad, M. Balazinska, J.-h. Hwang, W. Lindner, A. S. Maskey, A. Rasin, E. Ryvkina, N. Tatbul, Y. Xing, and S. Zdonik, "Aurora: a new model and architecture for data stream management," *The VLDB Journal*, vol. 12, no. 2, pp. 120–139, 2003.

[20] D. P. Arvind, A. Arasu, B. Babcock, S. Babu, M. Datar, K. Ito, I. Nishizawa, J. Rosenstein, and J. Widom, "Stream: The Stanford stream data manager," *IEEE Data Engineering Bulletin*, vol. 26, pp. 19–26, 2003.

[21] H. Schweppe, A. Z. Member, and D. Grill, "Flexible On-Board Stream Processing for Automotive Sensor Data," *IEEE Trans. Ind. Informat.*, vol. 6, no. 1, pp. 81–92, 2010.

[22] S. Geisler, C. Quix, S. Schiffe, and M. Jarke, "An evaluation framework for traffic information systems based on data stream," *Transportation Research Part C*, vol. 23, pp. 29–55, 2012.

[23] S. Geisler, S. Weber, and C. Quix, "Ontology-based Data Quality Framework for Data Stream Applications," in *Proc. 16th International Conference on Information Quality*, 2011.

[24] S. Schmidt, T. Legler, D. Schaller, and W. Lehner, "Real-Time Scheduling for Data Stream Management Systems," in *Proc. 17th Euromicro Conference on Real-Time Systems*. IEEE, 2005, pp. 167–176.

[25] Y. Wei, S. H. Son, and J. Stankovic, "RTSTREAM : Real-Time Query Processing for Data Streams," in *Proc. 9th IEEE International Symposium on Object and Component-Oriented Real-Time Distributed Computing*, 2006, pp. 141–150.

[26] M. A. Sharaf, P. K. Chrysanthis, and A. Labrinidis, "Preemptive Rate-based Operator Scheduling in a Data Stream Management System," in *Proc. 3rd ACS/IEEE International Conference on Computer Systems and Applications*, 2005, pp. 46–54.

[27] X. Li, Z. Jia, L. Ma, R. Zhang, and H. Wang, "Earliest Deadline Scheduling for Continuous Queries over Data Streams," *2009 International Conference on Embedded Software and Systems*, pp. 57–64, 2009.

[28] J. Gama and M. M. Gaber, Eds., *Learning from Data Streams*. Springer, 2010.

[29] W. Thies, M. Karczmarek, and S. Amarasinghe, "StreamIt : A Language for Streaming Applications," in *Proc 11th International Conference on Compiler Construction*, 2002, pp. 179–196.

[30] R. Müller, "Data stream processing on embedded devices," Ph.D. dissertation, ETH Zurich, 2010.

[31] C. Cranor, T. Johnson, and O. Spataschek, "Gigascope: a stream database for network applications," in *Proc. 2003 ACM SIGMOD international conference on Management of data*, 2003, pp. 647–651.

[32] H. Kargupta, R. Bhargava, K. Liu, M. Powers, P. Blair, S. Bushra, J. Dull, K. Sarkar, M. Klein, M. Vasa, and D. H. Vedas, "VEDAS: A mobile and distributed data stream mining system for real-time vehicle monitoring," in *Proc 4th SIAM International Conference on Data Mining*, 2004, pp. 300–311.

[33] H. Kargupta, K. Sarkar, and M. Gilligan., "Minefleet: an overview of a widely adopted distributed vehicle performance data mining system," in *Proc, 16th ACM SIGKDD international conference on Knowledge discovery and data mining*, 2010, pp. 37–46.

Ontology-based Change Propagation in Shareable Health Information Applications

Anny Kartika Sari
Dept. of Computer Science and Electronics
Gadjah Mada University
Yogyakarta, Indonesia

Wenny Rahayu
Dept. of Computer Science and
Computer Engineering
La Trobe University
Melbourne, Australia

***Abstract*—One of the most important challenges to be addressed when establishing an integrated smart health environment is the availability of shareable health data and knowledge which standardize the interoperability of components within the environment. Health ontologies are commonly utilized to enable interoperability between applications in such environment. However, the dynamic nature of health knowledge causes the need for frequent changes in health ontologies which then must be propagated to the relevant applications. A change propagation method that can efficiently streamline the change management from an ontology to all the applications which reference to it is proposed. A component called a *mapper* is used to manage the mapping between application terms and ontology concepts. The mapper is aimed to maintain the applications' access to the most up-to-date ontology concepts and to improve the semantic mapping between the application terms and the ontology concepts. Some rules are developed for the change propagation process. The evaluation of the method shows that the mapper can improve the mapping list in terms of: (i) correctness, by proposing a new mapping entry to substitute an existing one which is not valid anymore because ontology concept is deleted or changed; (ii) currency maintenance, by recommending a better mapping between an application term and a new ontology concept based on the similarity value between the term and the new concept.**

Keywords*—*health information system; ontology-based application; ontology evolution

I. Introduction

Current developments in the health domain require patients' data to be exchanged between information systems of different health care providers. Today, people are highly mobile, thus they can access health care treatment from different providers who can be geographically separated. Specialization in the health care domain also requires flexibility in the exchange of patient data. However, it requires the availability of shareable health data and knowledge which standardize the interactions of components within the environment.

Semantic interactions in the environment can be standardized using health ontologies. Several health ontologies such as SNOMED CT (Systematized Nomenclature of Medicine–Clinical Terms), UMLS (Unified Medical Language System) and LOINC (Logical Observation Identifiers Names) have been established to achieve semantic interoperability between different health providers in the environment. While an ontology-based health information system application must refer to the most current ontology, health ontologies constantly change due to changes in the knowledge of the health domain. The frequent changes in health ontologies may become a problem in the effort to ensure the applications refer to the most up-to-date ontology concepts.

In the notion of ontology evolution, there is a phase which is related to the effort of maintaining the currentness of the ontology-based applications with regard to the ontology to which they refer. This phase is named *ontology change propagation*. The goal of this phase is to bring the changes of the ontology to the depending artifacts such as other ontologies or the applications based on it. Since ontologies and applications have different characteristics, the ontology change propagation process is classified into two types: *ontology-to-ontology* change propagation and *ontology-to-application* change propagation. In this work, the focus is on the change propagation from a base ontology to the applications.

The complexity of change propagation process depends on how the ontology bound to the applications. The process will be easier for applications in which the concepts in the ontology are not tightly bound to them. This means that the ontology components are not hard-coded/embedded in the applications. Such applications only access the ontology 'on the fly', that is, the applications only need it when they are executing a process. For instance, in an ontology-based decision support system, an ontology is needed during the reasoning process to support decision making. In this type of application, there is no continuous direct binding between ontology concepts and application terms. The applications need the ontology as a whole, not only particular components. Once the ontology changes, the applications can be immediately directed to the new ontology. The impact of the ontology changes to the applications is not significant because the ontology components are not hard-coded in the applications.

For applications where ontology components are embedded, different approaches should be used. In this type of application, ontology components are used continuously, and may be hard-coded in the applications. For instance, in applications which utilize ontology to achieve interoperability between different health information systems, there may be one-to-one mapping between each term in the application and an ontology concept. In this type of application, changes in ontology should be propagated to applications straight away so that the applications always refer to the current ontology and interoperability is maintained. However, direct propagation may affect the validity of data or cause inconsistency. Furthermore, sometimes the nature of the applications make

it impossible to make frequent changes to the applications because they may raise some technical issues. In this type of application, a change propagation process which does not directly affect the applications would be more appropriate. To the best of our knowledge, there is no existing work on ontology-to-application change propagation which considers this matter.

In this paper, a method to handle ontology changes in an ontology-based application is proposed. The main focus of this work is the ontology-to-application change propagation process, specifically from the ontology to the depending applications which are constantly bound to the ontology concepts. A component, referred to as the *mapper*, is responsible for managing the mapping list such that the application terms can always be bound to the current ontology concepts. The main advantage of the mapper is that the binding between the application terms and the ontology concepts can be done outside the application so that the ontology changes can be handled without the need to modify the application. Some rules are developed as guidelines for the mapper to perform its task. The mapper has two important roles. Firstly, it can propose a new mapping entry to substitute an existing one which is not current due to a deletion or change of the ontology concept listed in the entry. Secondly, the mapper can propose a better mapping of an application term because a new concept in the ontology is found to be more semantically similar to the term than the existing concept previously bound to the term. In this way, the mapping list can be kept up-to-date, while its quality is improved.

The rest of the paper is structured as follows. Section 2 outlines related work. Section 3 presents the connection mechanism between ontology and the applications. The main focus of this work is discussed in Section 4, which explains the method in propagating the changes from the ontology to the applications and managing the mapping list when the ontology changes. Section 5 discusses the evaluation of the method. Section 6 concludes the work.

II. Related Work

Previous work on the management of ontology evolution has been proposed. Some of the important frameworks are CONCORDIA ([1], [2]), CREAM ([3]), OntoView ([4]), KAON ([5], [6]), CHAO ([7]), Evolva ([8], [9]), GOMMA ([10]), COnto-Diff ([11]) and CHO (Change History Ontology) ([12], [13], [14]). However, only a few of the frameworks address the change propagation phase, such as [2], [6], [7], [15], [16] and [17]. most of them only discusses the ontology-to-ontology change propagation method, not the ontology-to-application method. Our previous work in [18] and [19] also discusses an ontology-to-ontology approach with the focus on the change propagation from a base ontology to a sub-ontology. While the approach in [18] is only based on the change operations provided by the release of the health ontologies, the method proposed in [19] considers the semantic of the change operations.

The work on ontology evolution in ontology-based applications which is related to ontology-to-application change propagation method is summarized in Table I. The proposed approach differs from the existing work summarized in the table in at least two issues. Firstly, the focus is on health ontologies which have some specific characteristics as follow: (i) they have been standardized; (ii) they change very frequently; (iii) their size is very large, and; (iv) the domain is very critical. The ontologies which have become the focus of the existing work do not have these characteristics. Secondly, the main interest of the change propagation in this work is the change propagation to the applications which use the concepts constantly in the direct binding between the concepts and the application terms. To the best of our knowledge, there has not been any existing work which focuses on this issue. In most of the existing work on the ontology-to-application change propagation, including the work listed in I, the applications utilize ontologies in only two ways: instances and queries. In [24], an ontology-based method to handle terminology changes is proposed. The work focuses on International Classification of Diseases (ICD-9-CM) terminology, which is one of the standardized terminologies commonly used in medical area. However, the work does not consider the change propagation process from the terminology to the applications.

III. The Connection between Ontology and Applications

Figure 1 shows a description of the framework used in distributed health provider systems. It consists of a main ontology, a main ontology manager and several different health provider systems, each of which contains an ontology suitable for the system which is referred to as the *referred ontology*. Each health provider system also contains some health information system applications. The ontology manager is the key component in the framework. It manages the ontologies by doing two tasks: 1) keeping the ontologies up-to-date by propagating the changes which occur in the main ontology to the relevant referred ontologies and; 2) giving notification to each health provider system whenever there are changes in its referred ontology as a consequence of the changes in the main ontology. In this paper, only the second task is discussed because it is related to the mapping mechanism between ontology concepts and application terms.

As previously mentioned, continuous access to ontology concepts by the applications will include direct reference or mapping between the ontology concepts and the application terms. To make the references more well-structured, a *mapping list* is used. It includes the mapping between the terms used in the applications and the concepts included in the referred ontology. A component referred to as the *mapper* is also used to manage the mapping list and to handle the changes when the ontology evolves. Since the mapper and the mapping list are not part of the applications, the ontology components do not need to be hard-coded in the applications. By separating the mapping list and the mapper from the applications, management of the mapping when the ontology changes will be easier. The application does not need to do the adjustment every time the ontology changes because the mapping in the mapping list has been adjusted by the mapper.

Figure 2 shows the detail of the health provider system components and the processes which occur when the referred ontology changes. The tasks of the main ontology manager are to provide the referred ontology with the ontology components and to notify the mapper when there is an ontology change. It

TABLE I: Some existing works on ontology-to-application change propagation

Article	Ontology	Goal	Application
[20]	Legal Ontology	To discover inconsistencies in a semantic web service description, whose repairing improves the agreement of the ontology with the business rules	e-gov change management system
[3], [5]	not specified	To enable consistency in the annotations of knowledge sources in the case of changes in the domain ontology	CREAM, a semantic annotation framework
[21]	The Ontology of Professional Judicial Knowledge (OPJK)	To provide ontology managers and users with a tool that helps to detect effects of changes in ontologies and select versions based on their properties	MORE (Multi-version Ontology REasoner)
[22], [23]	CIDOC Conceptual Reference Model (CRM) ontology	Ontology-based system could provide continuous and unchanged services to the end-users	Applications based on CIDOC CRM ontology

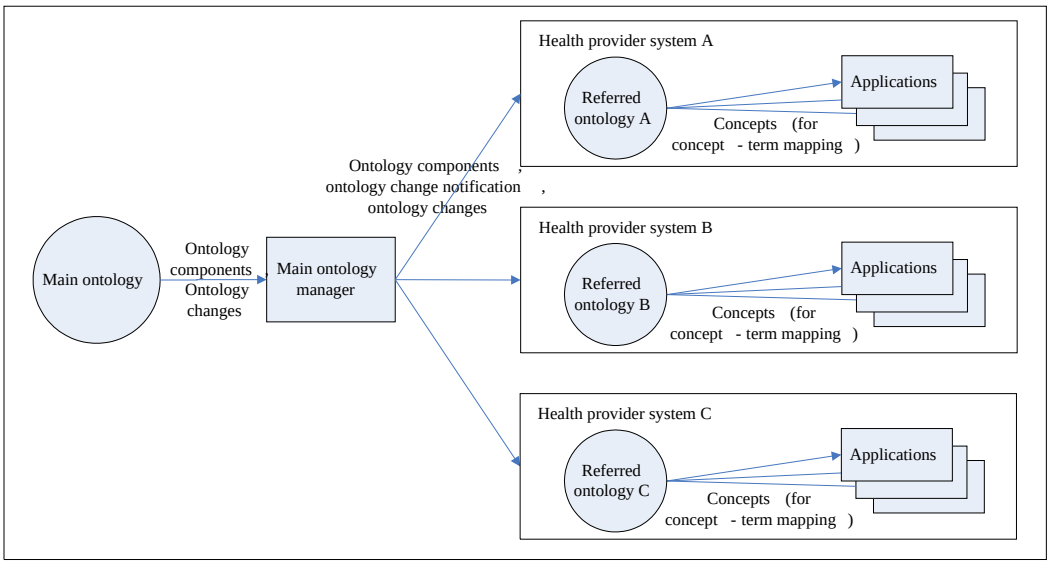


Fig. 1: Framework of the distributed health provider systems

also provides the mapper with a list of the change operations applied to the referred ontology. Based on the list of the change operations, as well as the term-concept mapping data provided by the mapping list, the mapper do the adjustment to the mapping list. The applications, or to be more specific the user/administrator of the applications, can accept or reject the adjustment.

At any given time, the health provider keeps the valid and up-to-date ontology, which is consistent with the main ontology, to be referred to by the applications. However, due to technical issues or data loss, sometimes there are applications which still refer to the concepts from the previous version of the ontology which are not included in the current version because they have been deleted or changed. To anticipate this situation, the old concepts are included as an extension of the valid ontology, which is referred to as the *extended concepts*. These extended concepts, together with the valid and up-to-date ontology, construct the referred ontology. Figure 3 shows the contents of a referred ontology. In the figure, 'Ontology' refers to the valid and up-to-date ontology. The extended concepts are not components of the valid ontology. Each of the extended concepts is annotated with the information of the concept in the valid referred ontology which is related to it, that is, the new concept which replaces or represents the extended concept. Moreover, information on the version of the main ontology in which it was originally included is also available. The formal definition of the annotation for the extended concept ce follows.

Definition 1. Annotation for ce
$A(ce) \equiv < c, v >$ *is the annotation for the extended concept ce where c is the concept in the current referred ontology related to ce and v is the version of the referred ontology in which ce was included.*

The mapping list contains several mapping entries, each of which connects a term used in the application to a concept in the referred ontology. The formal definition of the mapping list is as follows.

Definition 2. Mapping list
$L \equiv \{l_1, l_2, ..., l_n\}$ *is the mapping list with* $l_i \equiv < a_i, t_i, c_i, v_i, df_i, dt_i, s_i, r_i >$ *is the mapping entry.*

In the definition, L is the mapping list which is a set of mapping entries l_i. l is the mapping entry in the mapping list and has the following structure:

<application_id, term_id, concept_id, version, date-from, date-to, status, reason >

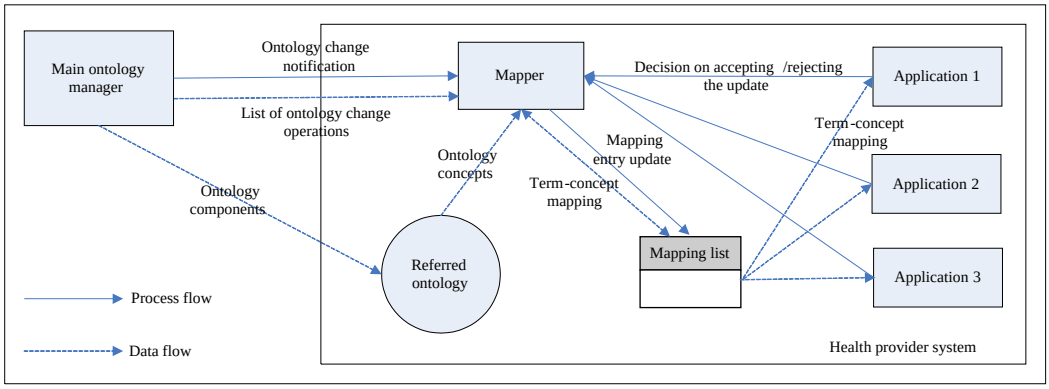


Fig. 2: The components inside a health provider system and the process and data flows when the main ontology changes.

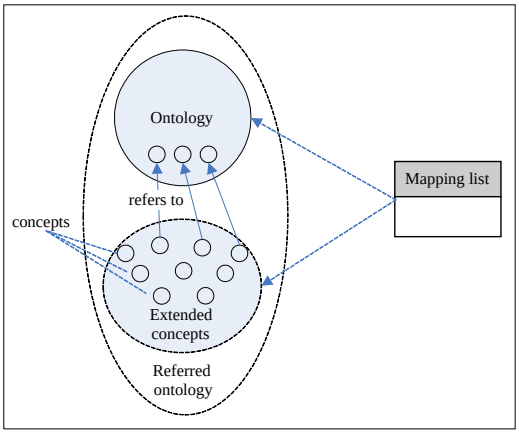


Fig. 3: The valid and up-to-date ontology and the extended concepts construct the referred ontology.

Application_id (a) is the ID of the application where the term is used. *Term_id* (t) is the ID of the term which refers to the ontology concept. *Concept_id* (c) is the ID of the concept in the referred ontology. Each concept which ID is currently used in the mapping list is referred to as a *referred concept*. *Version* refers to the most current version of the referred ontology where the concept is included. *Date-from* (df) is the date when the mapping is created. *Date-to* (dt) is the date when the mapping becomes obsolete due to ontology changes or application changes. *Status* (s) is the status of the mapping which can be: (i) 'valid', which means that the mapping is still used; (ii) 'invalid', which means that the mapping is not used anymore; (iii) 'new', which means that the mapping is produced by the mapper due to finding a new version of the referred ontology; or (iv) 'obsolete', which means that the mapping should be checked for its validity due to the change in the referred ontology. The *reason* (r) explains why a mapping entry is not valid anymore.

During the change propagation process, the mapping list can be changed. We define two operations which are related to the change of the mapping entries in the mapping list: *update* operation and *addition* operation. Deletion operation is not defined because the history of the mapping entries should be maintained in the mapping list. The formal definition of the update and addition operations is as follows.

Definition 3. Mapping entry update operation
$Given\ the\ mapping\ entry\ l_1 \equiv < a_1, t_1, c_1, v_1, df_1, dt_1, s_1, r_1 >$.
$UpdateMap(< a_1, t_1, c_1, v_1, df_1, dt_1, s_1, r_1 >, < a_2, t_2, c_2, v_2, df_2, dt_2, s_2, r_2 >) \equiv (a_2 \leftarrow a_1) \wedge (t_2 \leftarrow t_1) \wedge (c_2 \leftarrow c_1) \wedge (v_2 \leftarrow v_1) \wedge (df_2 \leftarrow df_1) \wedge (dt_2 \leftarrow dt_1) \wedge (s_2 \leftarrow s_1) \wedge (r_2 \leftarrow r_1)$

Definition 4. Mapping entry addition operation
$Given\ mapping\ list\ L \equiv \{l_1, l_2, ..., l_n\}$.
$AddMap(l_{n+1}, L) \equiv L \leftarrow L \cup \{l_{n+1}\}\ with\ l_{n+1} \equiv < a_{n+1}, t_{n+1}, c_{n+1}, v_{n+1}, df_{n+1}, dt_{n+1}, s_{n+1}, r_{n+1} >$.

Basically, the mapping entry update operation is used to update the field values of an existing mapping entry. One or more of the field values can be changed to the new and correct one. The mapping entry addition operation is applied when a new mapping entry needs to be added to the mapping list. The field values of the new mapping entry have been defined before the addition of the mapping entry.

IV. Change Propagation Process

The change operations in the ontology can be applied to all components of an ontology which can be concepts, relationships, description and description mappings. However, since generally the applications of the health provider system only use the concepts to be referred to, in this paper, only the changes related to concepts are considered. There are four types of concept change operations in the ontology:

1) $AddCon(c)$: adds a new concept c to the ontology. The concept can be a leaf or non-leaf concept.
2) $DelCon(c)$: deletes an existing concept c from the ontology. No new or existing concept is proposed to represent the meaning of c.
3) $ChangeCon(c_1, c_2)$: changes concept c_1 into c_2, and then c_1 is deleted from the ontology. This means that c_2 can be used to represent the meaning of c_1.
4) $MovCon(c)$: moves concept c to another branch in the ontology graph. Movement of a concept does not influence the entries in the mapping list because the concept still exists in the ontology.

Figure 4 shows the processes which occur in the mapper upon changes to the referred ontology. A list of the ontology change operations is provided by the main ontology manager. It contains the list of concepts which are deleted, added, moved or changed. When the main ontology has been changed, the main ontology manager performs the 'notify ontology changes' process. This process triggers the processes performed by the mapper. Three processes are performed by the mapper. The first process is the 'check mapping entries'. This process examines each mapping entry with a 'valid' status and checks whether the concept in that entry is listed in the list of the ontology change operations provided by the main ontology manager. The concepts which are not included in the mapper but are referred to by any extended concepts through their annotation are also examined. The second process is the 'update mapping list'. In this process, the mapper updates the mapping list according to the rules for the mapping list changes which will be discussed later. The updates, which can be editing or addition of entries, are proposed to the user for agreement. The last process is the 'update mapping list according to the user's decision'. In this process, the mapper updates the mapping list based on the user's decision on the proposed mapping list produced by the earlier process. We consider that a decision by the user (administrator, engineer or expert) is needed in this process because the user might not want to use the evolved version of the referred ontology or he may reject some of the new entries proposed by the mapper due to technical issues or other reasons. After the mapping list is changed in accordance with the user's decision, the referred ontology should be adjusted to suit the decision of the user. The adjustment is especially needed when the user chooses not to change the mapping entry, which specifically influences the extended concepts part of the referred ontology.

When the mapper receives notification from the main ontology manager that the referred ontology has been changed, it automatically searches for mapping entries with 'valid' or 'not used' status. If a concept in a mapping entry is also listed in the list of changes, or if a concept in the list of changes is referred to by an extended concept included in the mapper, the mapper will do the adjustment to the mapping entry by performing two actions:

1) It changes the status field from 'valid' to 'obsolete'. The reason field is also set to either 'deleted concept', 'changed concept', 'new child' or 'new parent', which is chosen according to the type of change operation corresponding to the entry.
2) As a replacement, it proposes new mapping entries with the same application_id and term_id values, but the value of the concept_id field is different, which is determined by several rules. The status field is set to 'new'.

Not all 'valid' mapping entries are affected by the ontology changes. Rules are defined to determine which mapping entries must be updated due to the corresponding change operation. The rules are built based on the type of change operation, as described above. In the following description, the impact of each change operation type is explained and underlies each rule. As previously mentioned, the $MoveConc$ operation does not influence the mapping list, thus the rule for the operation is not defined. Hence, there are only three rules which correspond to the three types of change operations.

1) Impact on the concept deletion operation ($Delcon(c_1)$)
When the concept_id field of a mapping entry refers to a concept to be deleted in the ontology, or when a concept to be deleted from the ontology is referred to by an extended concept contained in a valid mapping entry, the mapping entry must be updated. The value of the concept_id field must be changed to the ID of the *candidate concept*, i.e. the most similar child or parent concept of the deleted concept. The formal definition of the rule follows.

Rule 1: Propagation of $DelCon(c_1)$ operation
$\forall\ DelCon(c_1) \mid \exists\ l_1 \in L,\ l_1 \equiv < a_1,\ t_1,\ c_x,\ v_1,\ dd/mm/yyyy,\ null,\ 'valid',\ null > \ \wedge (c_x = c_1\ \vee\ c_x = ce \mid ce\ is\ an\ extended\ concept\ \wedge A(ce) \equiv < c_1,\ v_{ce} >) :$
$c_i\ is\ the\ candidate\ concept\ \rightarrow\ UpdateMap(l_1, < a_1,\ t_1,\ c_x,\ v_1,\ dd/mm/yyyy,\ null,\ 'obsolete',\ deletion' >) \wedge\ AddMap(< a_1,\ t_1,\ c_i,\ v_{current},\ null,\ null,\ 'new',\ 'null' >, L)$

The candidate concept is determined in the following way. The similarity between a concept c_1 and its parent c_2 is defined as $sim_{cp}(c_1, c_2)$ or $sim_{cp}(c_2, c_1)$, and the value is calculated using the child/parent pair similarity formula proposed in [25]. If c_1 is the deleted concept, while $c_2, ..., c_n$ are the set of parent or child concepts of c_1 in the previous referred ontology, $sim_{cp}(c_1, c_2), sim_{cp}(c_1, c_3), ..., sim_{cp}(c_1, c_n)$ are calculated. If there is c_i with $c_i \in c_2, ..., c_n$ where $sim_{cp}(c_1, c_i) = max\{sim_{cp}(c_1,\ c_2), ..., sim_{cp}(c_1, c_n)\}$, then c_i is referred to as the *candidate concept*. If there are $c_i, c_{i+1}, ..., c_{i+x} \in c_2, ..., c_n$ which have the same child/parent similarity value to c_1, MESH (Medical Subject Heading) vocabulary is used to find which of the names of $c_i, c_{i+1}, ..., c_{i+x}$ are in the same descriptor record to the name of c_1, in which the root words of the names of those concepts are used. If c_i is the only concept whose name is in the same descriptor record to the name of c_1 in MESH, c_i is the candidate concept. Otherwise, Jaro-Winkler distance is used to find the candidate concept whose name is the most similar to the name of c_1. Jaro-Winkler distance is adequate to find the candidate concept because using two level of assessment, i.e. child/parent similarity and MESH descriptor containment, basically the concepts have the same similarity value to the deleted concept. Hence, they are differentiated by their names.

2) Impact on the concept change operation ($ChangeCon(c_1, c_2)$)
When the concept_id field of a mapping entry refers to a concept to be changed in the referred ontology, that mapping entry must be updated. The concept_id field of the new proposed mapping entry must refer to c_2, which is the concept to which c_1 is changed in the ontology. The formal definition of this rule follows.

Rule 2: Propagation of $ChangeCon(c_1, c_2)$ operation
$\forall\ ChangeCon(c_1, c_2) \mid \exists\ l_1 \in L,\ l_1 \equiv < a_1,\ t_1,\ c_x,\ v_1,\ dd/mm/yyyy,\ null,\ 'valid',\ null > \ \wedge (c_x = c_1\ \vee\ c_x = ce \mid ce\ is\ an\ extended\ concept\ \wedge A(ce) \equiv < c_1,\ v_{ce} >) :$

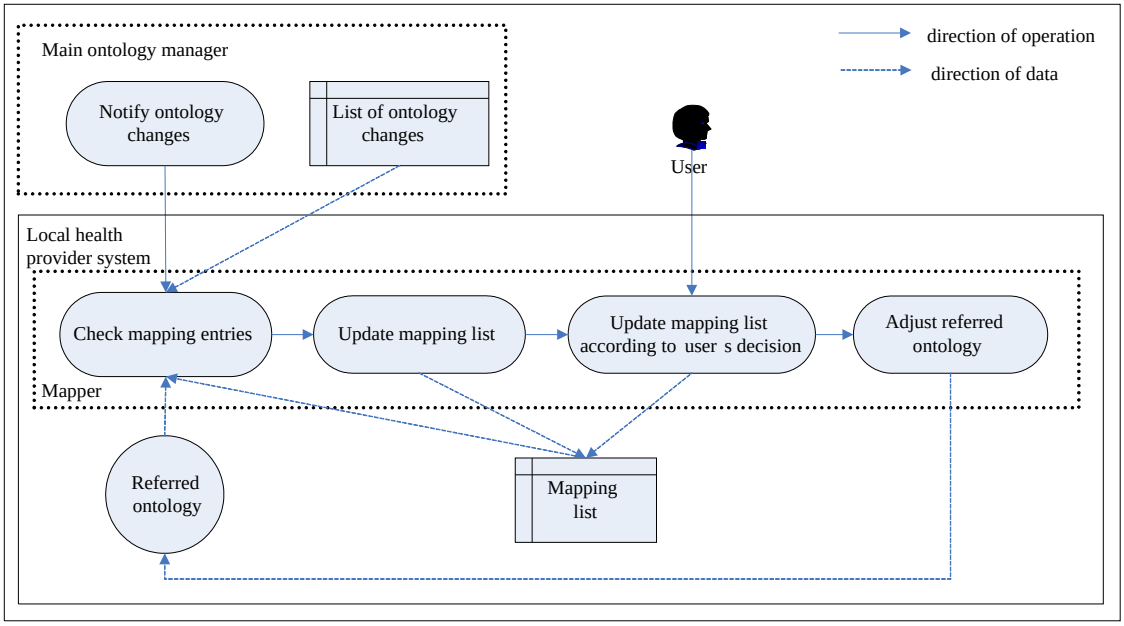


Fig. 4: The process of updating the mapping list when the ontology changes.

$UpdateMap(l_1, < a_1,\ t_1,\ c_x,\ v_1,\ dd/mm/yyyy,\ null,\ 'obsolete',\ 'change' >) \wedge\ AddMap(< a_1, t_1, c_2,\ v_{current},\ null,\ null,\ 'new',\ null >, L)$

3) Impact on the concept addition operation ($AddCon(c_2)$) When the concept_id field of a mapping entry refers to concept c_1 which happens to be the parent or child concept of the new added concept c_2, that mapping entry must be checked for the possibility of update. If t_1 is the value of the term_id in the mapping entry and $sim(t_1, c_2)$, i.e. the similarity value between t_1 and c_2, is higher than $sim(t_1, c_1)$, the mapping entry must be updated and a new mapping entry is proposed, otherwise nothing happens. The formal definition of the rule follows.

Rule 3: Propagation of $AddCon(c_2)$ operation
$\forall\ AddCon(c_2) \mid (\exists\ l_1 \in L,\ l_1 \equiv < a_1,\ t_1,\ c_x,\ v_1,\ dd/mm/yyyy,\ null,\ 'valid',\ null > \wedge\ (c_x = c_1\ \vee\ c_x = ce \mid ce\ is\ an\ extended\ concept\ \wedge A(ce) \equiv < c_1, v_{ce} >))$
$\wedge (c_1\ is\ the\ parent\ or\ child\ concept\ of\ c_2) :$
$sim(t_1, c_2) > sim(t_1, c_1) \rightarrow UpdateMap(l_1, < a_1, t_1, c_1, v_1, dd/mm/yyyy, null, 'obsolete', 'addition' >)$
$\wedge\ AddMap(< a_1, t_1, c_2,\ v_{current},\ null,\ null,\ 'new',\ null >, L)$

To evaluate the similarity values, MESH is used. If the name of c_2 is in the same descriptor record as the name of t_1 while it is not the case for the name of c_1, $sim(t_1, c_2) > sim(t_1, c_1)$. If it is the other way around, then $sim(t_1, c_1) > sim(t_1, c_2)$. The similarity value cannot be determined using MESH if one of these cases occurs: 1) the name of t_1 is not found in MESH descriptor records; 2) the names of c_1 and c_2 are in the same descriptor record as the name of t_1; 3) neither the name of c_1 nor c_2 is in the same descriptor record as the name of t_1. In these cases, Jaro-Winkler distance is used to find which concept name is more similar to the name of t_1.

After the mapper updates the mapping list based on the above rules, the user receives a notification that the mapping list has been updated due to the ontology change. This is included in the 'update mapping list by the user' process described in Figure 4. The user examines all mapping entries with status value 'new'. He should change the 'new' status to 'valid' if he agrees to the new mapping, or 'invalid' if he does not accept it. Following the changes by the user, the mapper adjusts the mapping list using the following guidelines:

- If the 'new' status is changed to 'valid', the status of the corresponding mapping entry with the same application ID and term ID with 'obsolete' status is changed to 'invalid' by the mapper and the date-to field is set to the current date. The date-from value of the new entry is set to the current date. Hence, the new mapping entry is used by the applications to replace the old one.
- On the other hand, if the 'new' status is changed to 'invalid', the status of the other mapping entry with the same application ID and term ID with 'obsolete' status is changed back to 'valid' by the mapper. The status field of the new entry is set to 'not used'. The 'not used' status of a mapping entry indicates that the entry has not been used by the applications since the changes to the referred ontology. In other words, the 'valid' mapping entry with the same application ID and term ID does not refer to the current ontology.

The mapping list adjustment due to the decision by the user may have an impact on the referred ontology, especially to the extended concepts. This happens when the user decides

not to use the new proposed mapping entry to replace the obsolete one in the case of deletion and change operations. If the concept ce referred to by the obsolete mapping entry has not been included in the extended concepts part of the referred ontology, it is included in that part with $A(ce) \equiv < c_1, v >$ where c_1 is the concept proposed by the mapper to replace ce in the mapping list and v is the version of the referred ontology where ce originated. If concept ce has been included in the extended concepts part, only the value of c in the annotation needs to be changed to c_1.

V. Evaluation and Discussion

The proposed method is evaluated by presenting a case study in which a health archetype is used as an example of the application. Following the case study is discussion on the efficiency of the use of the mapper compared to the common method and the maintenance of the semantic currency contained in the applications due to the new mapping entries proposed by the mapper.

A. Application of the Method to the Archetype Term Binding Process

In this section, an application of the proposed method is presented for managing the mapping between the application terms and the ontology concepts when changes occur in the referred ontology. We used an archetype as the representation of an application which refers to the referred ontology. An archetype is a model of specific domain knowledge, in this case, clinical knowledge. Each archetype describes a complete clinical knowledge concept such as 'diagnosis' or 'test result' [26]. For the referred ontology, a sub-ontology, which is derived from SNOMED CT as the main ontology, is developed. The method proposed in [27] is used to build the sub-ontology.

For this work, an archetype is created as a representation of an application. The archetype was named *tooth_care_summary*. The archetype was created by considering the concepts changed in SNOMED CT so that the types of change operations in the sub-ontology can be shown in the archetype. Thus, the archetype itself is a modified version to fulfill the above requirement. The definition of the *tooth_care_summary* archetype is shown in Figure 5, which is previewed using the Archetype Editor -arc built by Ocean Informatics. It contains 40 terms, each of which is bound to a SNOMED CT concept. Based on the concepts required by the binding, a sub-ontology is built. The sub-ontology was extracted from the 20110131 version of the International SNOMED CT edition. The sub-ontology contains 557 concepts and 645 relationships of SNOMED CT. Based on the archetype terms and the sub-ontology concepts, the mapper created a mapping list.

Figure 6 shows part of the initial binding between the archetype terms and the 20110131 version of SNOMED CT concepts viewed by the Archetype Editor. The node column presents the term names, the code column refers to the bound SNOMED CT concepts, while the release column shows the version of SNOMED CT in which each of the concepts is initially included. Note that this is not the actual mapping list.

After the 20110131 version, SNOMED CT has been changed and the newer version is the 20110731 SNOMED CT. To accommodate the changes, the sub-ontology has been changed as well. Some change operations were performed. The mapper checked the mapping entries to see if the change operations were affected by the change operations. Apparently, there were seven operations which were related to the mapping entries. The effect of each change operation to the related mapping entry is summarised in Table II. In Figure 6, the highlighted rows contain the concepts which are affected by the change operations.

The highlighted rows in Figure 7 are the binding between archetype terms and the SNOMED CT concepts which are affected by the sub-ontology changes. It can be seen that the rows contain different concepts ids and release versions of SNOMED CT from the concept ids and release versions contained in Figure 6.

B. Discussion

There are several issues related to the proposed method which will be discussed in this section. These issues are elaborated as follows.

The use of sub-ontologies to increase efficiency in change propagation process

In the proposed method, applications refer to sub-ontologies instead of the base ontology. The number of concepts in a sub-ontology is smaller than the number of concepts in a base ontology. When the base ontology changes, only the relevant changes are propagated to each of the sub-ontologies. In this way, the number of changes in each of the sub-ontologies is also smaller than the number of changes in the base ontology.

After updating its components according to the changes propagated by the base ontology, a sub-ontology then propagates these changes to the applications referring to it. This process is performed by the mapper by updating its mapping list according to the rules of sub-ontology to application change propagation. Since the number of change operations in the sub-ontology is smaller than the number of change operations in the base ontology, the mapper does not need much time to examine all the change operations which occurred in the sub-ontology. The time needed to examine the change operations will be longer if the sub-ontology does not exist, which implies that the mapper must look up all the change operations in the base ontology. In the case of SNOMED CT used in the evaluation of the method, the number of changes which occurred to Version 20110131, which is included in Version 20110731, is 8,697 operations. It will take time for the mapper to check whether each of the operations falls into one of the rules for propagating the changes to applications.

Benefits of the use of the mapper

There are several advantages of the use of the mapper in the change propagation process to applications as follows.

- The mapper maintains the history of the application term references to ontology concepts.
  The mapper never deletes its entries. Invalid (not used) entries are only marked by the 'invalid' value of the

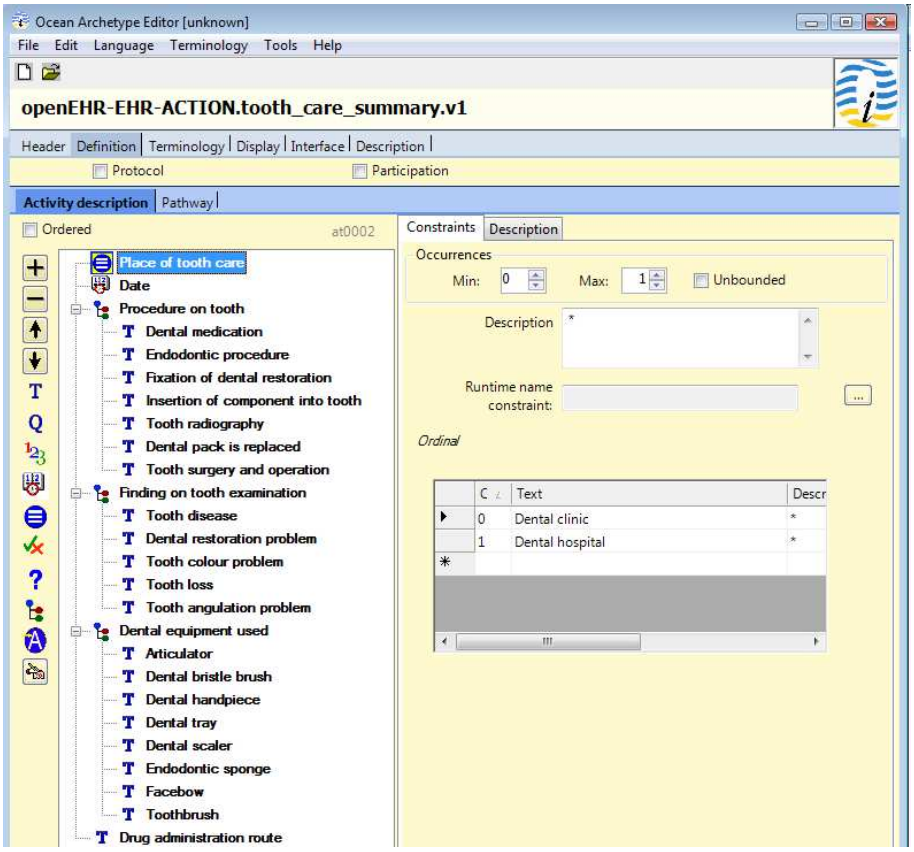


Fig. 5: The definition of *tooth_care_summary* archetype.

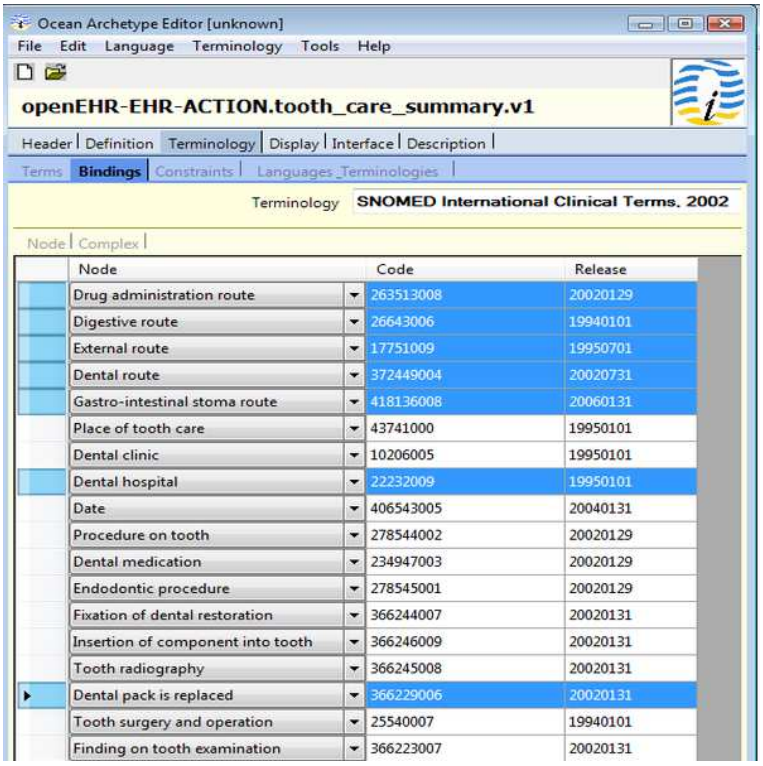


Fig. 6: The binding between the terms in the archetype and the SNOMED CT concepts before the sub-ontology changes.

TABLE II: Change operations found in the sub-ontology which are related to the mapping entries

Change operation in sub-ontology	Related mapping entry and the change					
	Term id	Term name	Concept id of the existing entry	Decision by the mapper	Reason	Concept id of the new entry
MergeCon (263513008, 410675002)	at0037	Drug admin-istration route	263513008	Status is changed to obsolete, new entry is proposed	Concept is merged	410675002
InsertCon (447964005)	at0038	Digestive route	26643006	Status is changed to obsolete, new entry is proposed	The term name is more similar to the name of the new concept	447964005
DelCon (17751009)	at0039	External route	17751009	Status is changed to obsolete, new entry is proposed	The deleted concept has only 1 parent concept (284009009), but no child concept	284009009
MoveCon (372449004)	at0040	Dental route	372449004	No need to update	The moved concept is still included in the sub-ontology	-
InsertCon (372454008)	at0041	Gastro-intestinal stoma route	418136008	No need to update	The term name is not more similar to the name of the inserted concept	-
AddLeaf (448399001)	at0004	Dental hospital	22232009	Status is changed to obsolete, new entry is proposed	The term name is more similar to the name of the new leaf concept	448399001
AddLeaf (447896001)	at0012	Dental pack is replaced	234718000	Status is changed to obsolete, new entry is proposed	The term name is more similar to the name of the new leaf concept	447896001

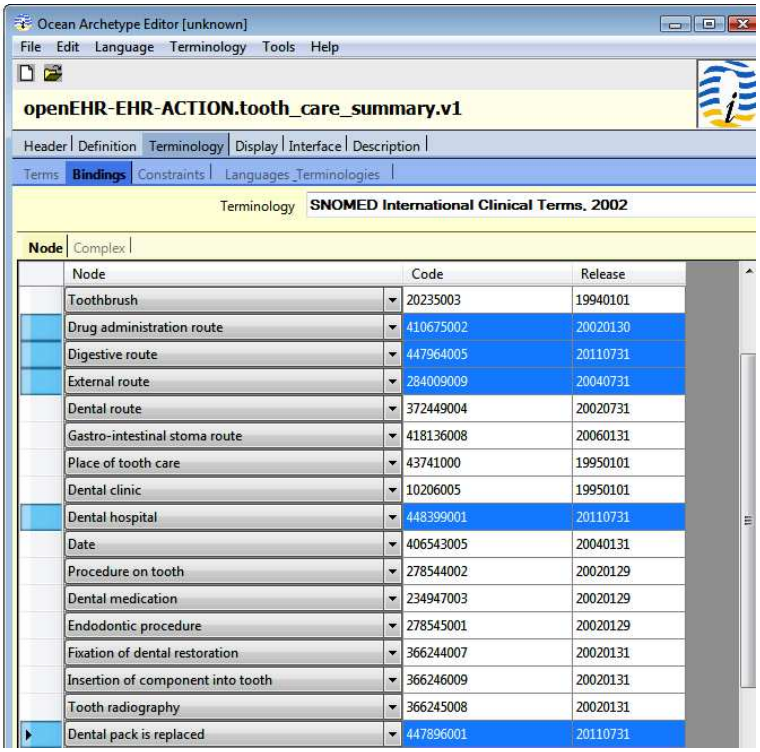


Fig. 7: The binding between the terms in the archetype and the SNOMED CT concepts after the sub-ontology changes.

status field. In this way, the history of references of application terms can be preserved. All references to a particular application term have the same values of *application_id* and *term_id* fields. Each invalid reference has a value for its *reason* field which records the reason why the reference is not used. Furthermore, the user can also check if there are application terms which are out-of-date and may need to be changed. The mapping entries with the most current value of the *version* field and 'not used' value of *reason* show that the corresponding application terms still refer to older ontology concepts.

- The mapper enables the semi-automatic update process of the mapping list.
  A fully manual update is error prone and obviously takes a longer time. In the sub-ontology used in the evaluation, there are 40 terms to be mapped to the sub-ontology concepts. In the case of a manual process, when the sub-ontology changes, the user must check each term to see whether it refers to a deleted sub-ontology concept or not. Since there are 557 concepts in the sub-ontology, the manual checking process is hardly feasible. The mapper can do the checking faster because it has the mapping list which includes the concept ids. Hence, the mapper only needs to find the concept ids which are included in the list of sub-ontology changes given by the central sub-ontology manager.

- The mapper facilitates the management of the mapping between the application terms and the sub-ontology concepts such that the sub-ontology changes can be handled without the need to modify the applications.
  This can be done because the mapping is managed outside the applications. The separation between the mapping mechanism and the applications is important because, unlike an ontology which can be changed automatically, the decision to change an application cannot be made instantly because it may affect the existing data and raise some technical issues related to the application development.

- The mapper can propose a better mapping of the application terms
  In the example, there are three terms which are mapped to the new concepts which are semantically more similar to the terms. In the previous mapping, the terms are mapped to the less similar concepts because they are the best choice in the previous sub-ontology. The new concepts added to the sub-ontology apparently have better similarity to the terms. Again, a manual examination takes time because there are 40 terms to be checked. The mapper can quickly propose the new mapping based on the rule of mapping list changes due to an addition of a leaf concept or an insertion of a concept. This will improve the quality of the mapping between the archetype terms and the ontology concepts. Otherwise, the application terms will maintain their references to less similar concepts, while there are actually concepts which have better similarity to them. At this moment, a mapping is improved by changing the referred concept to its new child or parent concept only in the changed sub-ontology. In the future, improvements might be made by changing the referred concept to any new concept in the sub-ontology or even any new concept in the base ontology. However, a discussion on this issue is beyond this thesis.

Validity of the mapping entries with regard to the current sub-ontology

A concept deleted from the sub-ontology indicates that the concept is not available in the current sub-ontology. In some ontologies such as SNOMED CT, a concept merged to another one is also considered a deleted concept. A reference to a non-existent concept is obviously not valid, and hence, should not exist in the mapping list. In the sub-ontology, the concept to be deleted is concept 17751009, while the concept to be merged is concept 263513008. In Figure 6, those two concepts are listed in the binding list between the archetype terms and the sub-ontology concepts. However, in Figure 7, these two concepts are not listed in the binding list. This is correct since the two concepts do not exist in the current sub-ontology. The terms previously bound to the two concepts are now binding to other concepts which exist in the sub-ontology. This shows that the proposed method is able to keep the application terms referring to the valid concepts in the current sub-ontology.

Possibility of application change due to sub-ontology change

The changes to the sub-ontology suggest that the knowledge has changed as well. For an application which has a very high requirement for knowledge update, the changes in the sub-ontology can be interpreted as an indication that the application needs to be updated. For instance, if a concept referred to by an application term is deleted from the sub-ontology, it may be the case that the term should be deleted from the application due to its obsoleteness. The mapper can give notification to the applications with regard to the changes, and it is the decision of the applications to update the terms included in them. If the terms are updated, the sub-ontology must be updated too because the selected concepts which are referred to by the application terms might be changed. This leads to another process of sub-ontology changes.

VI. Conclusion

In this paper, a change propagation mechanism from an ontology to the depending application has been proposed. A mechanism which is able to manage the continuous access of the applications to the ontology they refer to is proposed. Using the mechanism, the applications keep referring to the up-to-date ontology even when the ontology changes.The heart of the mechanism is the component called the mapper, which includes a mapping list. The mapping list contains mapping entries, each of which represent the mapping between an application term and an ontology concept. The task of the mapper is to manage the mapping list in the event of ontology changes so that the reference to a non-existing concept is avoided and the quality of the mapping entries can be improved by proposing new mapping entries which contain more relevant pairs of application terms and ontology concepts. Some rules,

which are based on the ontology change operations, have been created for the mapper to update the mapping list.

The proposed method has been applied to an application together with its referred ontology. An archetype is used to represent an application, while an ontology has been developed based on the SNOMED CT ontology and the archetype terms. It is shown that the mapper offers more efficient change propagation than the commonly used method in terms of the number of change operations which should be propagated to the applications. The use of the mapper also enables semi-automatic updating to the mapping list which is obviously faster than manual inspection. Moreover, the rules used by the mapper can maintain the semantic currency of the mapping because the mapper can propose a new mapping entry in which the application term is semantically more similar to the new concept than the previous one.

For future work, the application of the method in other domains, such as bioinformatics and Internet of Things, can be observed. Similar to the application of the method in health domain, the application to other domains requires the availability of a standardised ontology, the distributed environment nature, and the reference of application terms to ontology concepts. Technical aspects of the application of the approach may also be interesting for future work. The performance, reliability and scalability of the deployment of the approach in distributed environment needs to be examined.

REFERENCES

[1] D. E. Oliver and Y. Shahar, "Change management of shared and local health-care terminologies," Methods of Information Medicine, vol. 39, pp. 278–290, 2000.

[2] D. E. Oliver, Y. Shahar, E. H. Shortliffe, and M. A. Musen, "Representation of change in controlled medical terminologies," Artificial Intelligence in Medicine, vol. 15, pp. 53–76, 1999.

[3] N. Stojanovic, L. Stojanovic, and S. Handschuh, "Evolution in the ontology-based knowledge management systems," in Proceedings of the 10th European Conference on Information Systems, Gdansk, Poland, 2002, pp. 840–850.

[4] M. Klein, D. Fensel, A. Kiryakov, and D. Ognyanov, "Ontology versioning and change detection on the web," in Proceedings of the 13th International Conference on Knowledge Engineering and Knowledge Management (EKAW02), LNAI 2473, A. Gomez-Perez and V. Benjamins, Eds. Springer-Verlag, Berlin, Heidelberg, 2002, pp. 197–212.

[5] L. Stojanovic, A. Maedche, B. Motik, and N. Stojanovic, "User-driven ontology evolution management," in Proceedings of the 13th International Conference on Knowledge Engineering and Knowledge Management. Ontologies and the Semantic Web, ser. EKAW '02. London, UK: Springer-Verlag, 2002, pp. 285–300.

[6] A. Maedche, B. Motik, L. Stojanovic, R. Studer, and R. Volz, "Managing multiple ontologies and ontology evolution in ontologging," in Intelligent Information Processing. Kluwer, 2002, pp. 51–63.

[7] N. F. Noy, A. Chugh, W. Liu, and M. A. Musen, "A framework for ontology evolution in collaborative environments," in Proceedings of the 5th International Semantic Web Conference, ser. LNCS Volume 4273. Springer, 2006, pp. 544–558.

[8] F. Zablith, "Dynamic ontology evolution," in International Semantic Web Conference (ISWC) Doctoral Consortium, Karlsruhe, Germany, 2008.

[9] F. Zablith, M. Sabou, M. d'Aquin, and E. Motta, "Ontology evolution with evolva," in Proceedings of the 6th European Semantic Web Conference (ESWC) LNCS 5554. Springer-Verlag, Berlin, Heidelberg, 2009, pp. 908–912.

[10] T. Kirsten, A. Gross, M. Hartung, and R. Erhard, "Gomma: a component-based infrastructure for managing and analyzing life science ontologies and their evolution," Journal of Biomedical Semantics, vol. 2, 2011.

[11] M. Hartung, A. Grob, and E. Rahm, "Conto-diff: generation of complex evolution mappings for life science ontologies," Journal of Biomedical Informatics, vol. 46 (2013), pp. 15–32, 2013. [Online]. Available: http://dx.doi.org/10.1016/j.jbi.2012.04.009

[12] A. M. Khattak, K. Latif, S. Khan, and N. Ahmed, "Managing change history in web ontologies," in Proceedings of the Fourth International Conference on Semantics, Knowledge and Grid, China, 2008.

[13] A. M. Khattak, Z. Pervez, S. Lee, and Y.-K. Lee, "After effects of ontology evolution," in Proceedings of the 5th International Conference on Future Information Technology (FutureTech), 2010, pp. 1 – 6.

[14] A. M. Khattak, K. Latif, and S. Lee, "Change management in evolving web ontologies," Tsinghua Science and Technology, vol. 37 (2013), pp. 1–16, 2013. [Online]. Available: http://dx.doi.org/10.1016/j.knosys.2012.05.005

[15] J. Mapoles, C. Smith, J. Cook, and B. Levy, "Strategies for updating terminology mappings and subsets using snomed ct," in Proceedings of the 3rd international conference on Knowledge Representation in Medicine, R. Cornet and K. Spackman, Eds., 2008.

[16] A. Shaban-Nejad and V. Haarslev, "Bio-medical ontologies maintenance and change management," in Biomedical Data and Applications, ser. Studies in Computational Intelligence Volume 224, A. Sidhu and T. Dillon, Eds. Springer, 2009, pp. 143–168.

[17] R. Palma, O. Corcho, A. Gmez-Prez, and P. Haase, "A holistic approach to collaborative ontology development based on change management," Web Semantics: Science, Services and Agents on the World Wide Web, vol. 9, pp. 299–314, September 2011.

[18] A. K. Sari and W. Rahayu, "A methodology for change propagation in health ontology," in Proceedings of the 15th Pacific Asia Conference on Information Systems, Brisbane, Australia, 2011.

[19] A. K. Sari, W. Rahayu, and M. Bhatt, "An approach for sub-ontology evolution in a distributed health care enterprise," Information Systems, vol. 38, pp. 727–744, July 2013.

[20] L. Stojanovic, A. Abecker, N. Stojanovic, and R. Studer, "On managing changes in the ontology-based e-government," in Proceedings of the 3rd International Conference on Ontologies, Databases and Application of Semantics (ODBASE 2004), ser. LNCS Volume 3291. Springer Verlag, 2004, pp. 1080–1097.

[21] Z. Huang and H. Stuckenschmidt, "Reasoning with multi-version ontologies: A temporal logic approach," in Proceedings of the 4th International Semantic Web Conference (ISWC), 2005, pp. 398–412.

[22] Y. Liang, H. Alani, D. Dupplaw, and N. Shadbolt, "An approach to cope with ontology changes for ontology-based applications," in Second Advanced Knowledge Technologies DTA Symposium, Aberdeen, Scotland, 2006.

[23] Y. Liang, H. Alani, and N. Shadbolt, "Changing ontology breaks the queries," in Doctoral Symposium of The 5th International Semantic Web Conference, ser. LNCS Volume 4273. Athens, GA, U.S.A: Springer-Verlag, 2006, pp. 982–985.

[24] A. C. Yu and J. J. Cimino, "A comparison of two methods for retrieving icd-9-cm data: The effect of using an ontology-based method for handling terminology changes," Journal of Biomedical Informatics, vol. 44, pp. 289–298, 2011.

[25] H. Gu, J. Geller, L.-m. Liu, and M. Halper, "Using a similarity measurement to partition a vocabulary of medical concepts," in Proceedings of the 10th International Conference on Database and Expert Systems Applications (DEXA '99). London, UK: Springer-Verlag, 1999, pp. 712–723.

[26] H. Leslie and S. Heard, "Archetypes 101," in HIC 2006, J. Westbrook and J. Callen, Eds. Sydney: Health Informatics Society of Australia Ltd (HISA), 2006.

[27] A. K. Sari, W. Rahayu, and M. Bhatt, "Archetype sub-ontology: Improving constraint-based clinical knowledge model in electronic health records," Knowledge Based Systems, vol. 26, pp. 75–85, February 2012.

Similarity Calculation Method of Chinese Short Text Based on Semantic Feature Space

Liqiang Pan, Pu Zhang, Anping Xiong
College of computer science and technology
Chongqing University of Posts and Telecommunications
Chongqing, China

***Abstract*—In order to improve the accuracy of short text similarity calculation, this paper presents the idea that use the history of short text messages to construct semantic feature space, then use the vector in semantic feature space to represent short text and do semantic extension, and finally calculate the short text similarity of corresponding vector in the semantic feature space. This method can represent the semantic information of short text message thoroughly so as to improve the accuracy of similarity calculation. We selected a large number of problem test sets for experiments. The results show that the method we proposed is reasonable and effective.**

Keywords—short text; semantic feature space; similarity; semantic similarity

I. INTRODUCTION

With the wide application of short text similarity calculation method in information retrieval, question-answering system, text mining and other natural language processing fields, the research and improvement on the calculation method of short text similarity has become an important research hotspot. The research finds that there are many differences between the calculation methods of short text similarity and document similarity. As the document contains large amount of word information, most of the similarity calculation method is based on word statistical method. However, the short text contains little word information, maybe even only one word. It is not sufficient to judge the similarity between the short texts accurately only using the information of the short text itself. Therefore, in order to improve the calculation accuracy of short text similarity, we need to solve two key problems. The first problem is how to fully expressed and reflected short text information? The information includes word frequency, word meaning, etc. The second problem is how to calculate the similarity between the short texts? In order to solve these two problems, this paper presents the calculation method of Chinese short text semantic similarity based on the semantic feature space. This method represent the semantic information of short text message thoroughly so as to improve the accuracy of similarity calculation. We selected a large number of problem test sets for experiments. The results show that the method we proposed is reasonable and effective.

II. CONSTRUCTION METHOD OF SEMANTIC FEATURE SPACE

We take the intelligent-service system as the research background. The main short texts in the system are advisory information (namely interrogative sentences) and response short texts. In the intelligent service system, there are many users asking for advices every day, which inevitably produces massive consultation information. We can use these historical advisory information, namely short text sets to construct the semantic feature space, and then build the model by using the new consultation of the users or questioning short text in the space, finally we can calculate the similarity between the new short text and historical short text. The semantic feature space has a similar construction process with the ordinary vector space, which also consists of two main steps: feature selection and feature dimension reduction.

A. The feature selection of the semantic feature space

As the short text contains few words and may even contains only one single word, this paper only uses the feature of first level instead of phrase level because the feature of phrase level is not conducive to fully represent short text.

The initial feature set of semantic feature space FS' is constructed like this: first, segment all the historical short text data set and remove stop words (stop words have no effect on the semantic expression of the sentence); then, remove function words and remain content words according to function word table. This is because the semantic meaning of short text is mainly conveyed by content words, while function words are mostly auxiliary words of mood and not carrying much semantic information. At last, the initial feature set FS' is obtained by aggregating all the content words of short text A_i like this:

$$FS' = A_1 \cup A_2 \cup A_3 \cdots A_n \quad (1)$$

B. The feature dimension reduction based on semantic clustering

Because of the complexity and diversity of Chinese word structure, the space dimensions of the initial feature set FS' are particularly high. The direct use of the initial feature set will inevitably increase the complexity of similarity calculation. Through the experimental analysis, it is found that there are many lexical items with the same semantic meaning in the initial feature set FS'. Therefore, we use the word similarity calculation method based on "hownet"[1] to cluster the feature lexical items with higher similarity in the initial feature set. The basic idea of this clustering method is: first aggregate the feature item with higher similarity as a cluster, then choose one feature lexical item optionally as the representative, finally constitute a set of

all the representative feature lexical items. This set is the feature set of the final semantic feature space FS. Semantic clustering greatly reduces the dimension of the semantic feature space and redundant features of the initial feature set FS', so as to improve the efficiency of calculation of text similarity. The semantic clustering method is a bottom-up clustering algorithm. The pseudo code of the algorithm is as follows:

1. Initialize each feature as cluster, the whole cluster set $C=\{c_1,c_2,\cdots,c_i,\cdots,c_n\}$, max=0;
2. For i=1 to n do
3. For j=i+1 to n do
4. Calculate the similarity between c_i and c_j, denoted as csim(i,j);
5. If csim(i,j) > max
6. Max = csim(i,j);
7. k_1 =i; k_2 =j;
8. End if
9. End do
10. End do
11. If max > λ
12. Merge c_{k_1} and c_{k_2};
13. Update index of each cluster;
14. n = n-1;
15. Go to (2);
16. Else stop;
17. End if

The value of λ is range from 0 and 1. The calulate algorithm csim(i,j) is shown as follows:

1. Initialize csim(i,j)=0;
2. For each feature fw_k in cluster c_i do
3. For each feature fw_l in cluster c_j do
4. csim(i,j) = csim(i,j) +sim(fw_k , fw_l);
5. End do
6. End do
7. csim(i,j) = csim(i,j) / ($|c_i|\times|c_j|$);

The csim(fw_k, fw_l) is a sematic similarity method based on "hownet". This thesis carries on the detailed introduction to the semantic similarity calculation method[1]. The $|c_i|$ and $|c_j|$ is the feature number of C_i and C_j.

The initial feature set becomes the feature set FS after semantic clustering, which is used to construct the semantic feature space. Each feature lexical term in the semantic feature space expresses specific semantic meaning and subject. The construction of the semantic feature space needs a lot of corpus training and aggregation calculation, but as long as the first training corpus is enough, the semantic feature space can be used directly later.

III. THE SIMILARITY CALCULATION METHOD OF CHINESE SHORT TEXT BASED ON SEMANTIC FEATURE SPACE

For an arbitrary short text C, after recognition of center word and word frequency statistics, we can map it to the semantic feature space mentioned in the previous paper and build the model of the text, then calculate the similarity between short texts in the semantic space. The specific methods are as follows:

First, in order to obtain the part of speech tagging word set T, for the short text C do segmentation using automatic segmentation system, pos tagging and remove stop word ased on the stoplist. Then statistic word frequency of tagging word set T,we can use the word frequency initialization vector V_c express short text:

$$V_C=(tf_1,tf_2\cdots tf_i\cdots tf_m) \tag{2}$$

In(2), m represents the number of words have distinct speech tagging of word set. tf_i is word i 's Frequency intagging word set T. Because of the short text word have less information, common words's and the central word's word frequency values are often in the same or is 1. In order to express the importance of the center word which reflects the importance meaning of the short text, Center word's word frequency is need to heavier its weights. First, using Tian Weidong's[2] center word recognition method to recognize center word set Z. Then, for each center word's word frequency multiply a weighting factor η. So obtain a new vector representation V_c of short text:

$$V_q=(tf_1w_1,tf_1w_2\cdots tf_iw_i\cdots tf_mw_m) \tag{3}$$

In(3), w_i represents the weight of the word, its value is η (indicating central word) or 1 (indicating non-central word).

Then, all the features word of semantic feature spacethe FS and all the words of intagging word set T construct a similarity matrix which is also known as text mapping matrix.

$$A=\begin{bmatrix} a_{11} & \cdots & a_{1n} \\ \vdots & \ddots & \vdots \\ a_{m1} & \cdots & a_{mn} \end{bmatrix} \tag{4}$$

The n represents the word number of semantic feature vector space FS. a_{ij} denotes the semantic similarity between the i word in intagging word set T and the j word in semantic feature spacethe FS. pan's method[1] to calculate the similarity of the two words.

After the text mapping matrix is constructed, short text semantic mapping vector V_C' can be obtained in semantic space. The method is short text word frequency vector V_C is multiplied by the mapping matrix A.

$$V_q'=V_q\times A$$

$$=(tf_1w_1,tf_2w_2,\cdots,tf_iw_i,\cdots,tf_mw_w)\times\begin{bmatrix} a_{11} & \cdots & a_{1n} \\ \vdots & \ddots & \vdots \\ a_{m1} & \cdots & a_{mn} \end{bmatrix}$$

$$= (\sum_{i=1}^{m} tf_i w_i a_{i1}, \sum_{i=1}^{m} tf_i w_i a_{i2}, \cdots, \sum_{i=1}^{m} tf_i w_i a_{im}) \quad (5)$$

Using the mapping matrix A, we put a m dimensional vector V_C converted into a n dimensional vector $V_C^{'}$,The m less than n. By this method, we can use more features represent the semantic information of short text, so as to improve the accuracy of calculating the similarity between the short text.

Therefore, for any two short text C_1 and C_2, modeling the short text by the above methods. The two short text are mapped to the semantic feature space, so as to obtain vector V_{C_1} and V_{C_2}.

Finally, using vector cosine value represents two short text similarity $Sim(C_1, C_2)$:

$$Sim(C_1, C_2) = \cos(V_{C_1}^{'}, V_{C_2}^{'}) = \frac{V_{C_1}^{'} \cdot V_{C_2}^{'}}{|V_{C_1}^{'}| \cdot |V_{C_2}^{'}|} \quad (6)$$

IV. The Design and Analysis of the Experiment

In this section, we mainly set parameters and validate the similarity calculation method of Chinese short text based on semantic feature space described above through several experiments. Four steps will be introduced in this section. They are experimental data, experimental evaluation method, experiment setup and tool and experimental results and analysis.

A. Experimental data

The experimental data are purchased through data [3]. High quality question and answer corpus data set of Q & A community in 2013 (including 3000000 question and answer, database format, XML) are recorded as the original data set. As this paper only judges the similarity between short texts of problem, we only need short texts of problem. Through the analysis of the original data set, we write the preprocess programs, and extract 10 categories of problems according to the classification label of XML format (2000 short texts of problem for each category, a total of 20000) in order to form experimental data set D.

B. Experimental evaluation method

- Evaluate the similarity calculation method using $F-Value$

Take similarity calculation results between the texts as the similarity measure of K-Means clustering algorithm, then evaluate the effectiveness of similarity method. Metric data $F-Value$ is a balance index[4] of combination precision ration and recall ratio in information retrieval. The metric data $F-Value$ allows us to test whether the short text is correctly classified into the corresponding categories after clustering and the text of expected category is included in the same category.

Set the number of short texts of category i is n_i, the number of short texts of cluster j is n_j, n_{ij} represents the number of short texts which belongs to the cluster j and the category i, and then the precision ratio $p(i, j)$ of cluster j, the recall ratio $R(i, j)$ of category i can be respectively defined as:

$$P(i, j) = \frac{n_{ij}}{n_j}, R(i, j) = \frac{n_{ij}}{n_i} \quad (7)$$

The corresponding $F-Value$ $F(i, j)$ is defined as

$$F(i, j) = \frac{2 \times P(i, j) \times R(i, j)}{P(i, j) + R(i, j)} \quad (8)$$

Thus,we can get the global $F-Value$ F:

$$F = \sum_i \frac{n_i}{n} \max_j (F(i, j)) \quad (9)$$

The n represents the number of short text in the entire data set. The same as ordinary clustering algorithm, the larger $F-Value$, the better the effect of clustering can be inferred from the similarity algorithm

- Evaluate the similarity calculation method using P@n

In the practical application of intelligent service system or automatic question-answering system, similarity calculation method mainly calculates by comparing every short text of historical problems and then responses by outputting the most similar answer of short text of historical problems. Therefore, we can use the P@n (Precision at n) as our experimental evaluation standard. P@n represents the probability or proportion of the occurrence of the correct result (historical short text and pending short text is similar indeed) in the top n results. For example: P@4=0.5 means that there are 2 short texts similar indeed to the pending short text in the first 4 similar short texts after the similar calculation of pending short text and historical short text which is in descending order according to the similarity of the historical short texts.

C. Experiment setup and tool

K-Means clustering algorithm uses open source tool lingPipe to realize.

D. Experimental results and analysis

In the construction of semantic feature space, semantic feature space dimension has a close relationship with parameter λ. Figure 1 shows the process that dimension varies with the parameter λ (including 3 experiments). In each experiment, all the content words in data set D construct the semantic feature space initially, then set the similarity threshold value λ, make dimensionality reduction of semantic feature space and compute corresponding spatial dimensions. Because of complexity and diversity of Chinese words structure, the original space dimension is particularly high which can achieve several thousand dimensions. After screening of content words, words with no semantic meaning can be removed and the space dimension can be reduced to about 6200. We use the similarity calculation method to cluster the features of semantic similarity in semantic feature space, so as to further reduce the space dimension. When λ is small, more semantic features cluster in one category, thus the number of categories will be less. As the number of space dimension and clustering is the same, the space dimension is lower. When λ increases

gradually, the situation is the opposite.

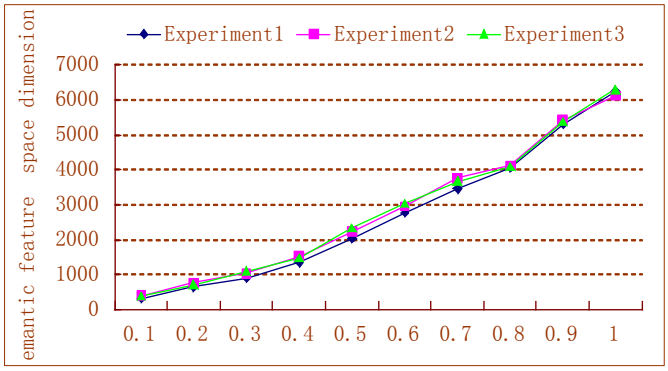


Fig. 1. semantic feature space dimension varies with the parameter λ

In order to get the best threshold value λ, we can observe the relationship between it and $F-Value$ in clustering algorithm, that is to say when $F-Value$ is the maximum (similarity effect is the best), value λ will be the best. Figure 2 is the variation diagram of threshold value λ and clustering algorithm, in which we can see that the optimal threshold value is between 0.4 and 0.6. Figure 2 only shows the results of three experiments. In every experiment, we calculate the similarity among data set D with different value λ using the similarity calculation method proposed in this paper, then take the results as the similarity measure of K-Means clustering algorithm, output corresponding $F-Value$ and form the graph. K-Means algorithm is implemented by using open source tool lingPipe. Value λ influences $F-Value$ of clustering algorithm by affecting space dimension. When the value λ is small, the space dimension is low, thus the semantic meaning of short text is not fully expressed, which leads to low $F-Value$ of clustering algorithm. When the value λ is larger, the space dimension is higher and a lot of invalid features and noises are introduced, thus the expression of the semantic meaning of the text is influenced, which results in lower $F-Value$ of clustering algorithm.

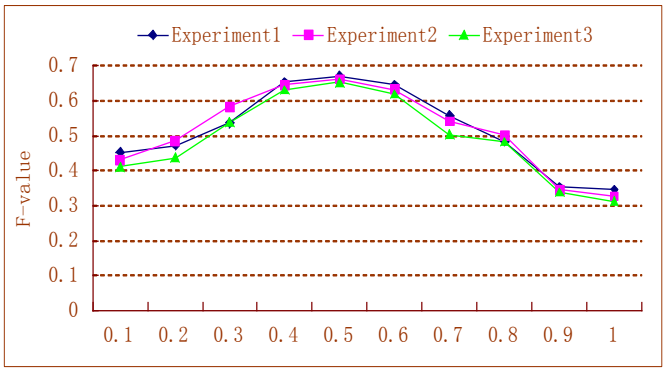


Fig. 2. $F-Value$ varies with the parameter λ

When building the model of the short text using semantic feature space, we add weight to the center word. The value of the weight will also affect the effect of similarity algorithm. Figure 3 is the relation graph of the center word weight η and $F-Value$. It shows the experimental results among many tests. In every experiment, the optimal λ is set to 0.52, the corresponding clustering algorithm $F-Value$ is output through changing the value of β the, thus the relation graph is formed. When $2 \le \eta \le 6$, $F-Value$ increases with the η, which shows that giving higher weights to the center word is helpful to improve the accuracy of the clustering, in other word, it is conducive to the similarity calculation. But with the increase of η, there is a downward trend of $F-Value$. The reason is the weight of the center word is so high that the function of other words is negligible and their semantic information is ignored. Therefore, the weight of the center word needs to be set to an appropriate value. Through repeated experimental analysis, the η should be set between 4~6.

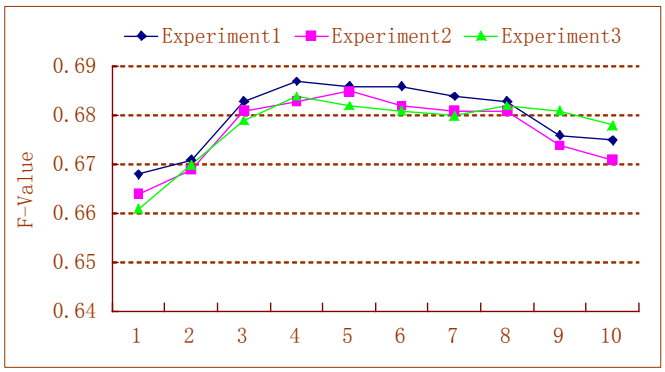


Fig. 3. $F-Value$ varies with the parameter η

We set the optimal parameters λ =0.52, η = 4 for the method in this paper. Then we conduct comparative tests with Huang Chenghui's text similarity measure method[5] which combines word semantic information and TF-IDF method and Song Wanpeng's question similarity calculation method[6] in question-answering system. Figure 4 shows that the similarity calculation method of Chinese short text based on semantic feature space can effectively improve the clustering effect, that is effectively judge the similarity between the short texts.

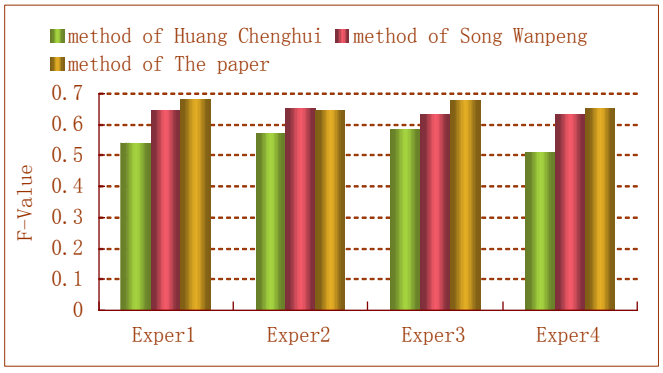


Fig. 4. Comparative experiment on Chinese short text similarity methods

To validate the effect of similarity calculation method in the actual application, we also use P@n to evaluate the similarity calculation method. P@n represents the probability or proportion of the occurrence of the correct result (historical short text and pending short text is similar indeed) in

the top n results. Table 1 presents the corresponding accuracy and also lists the accuracy of question similarity calculation method[11] in question-answering system of Song Wanpeng.

TABLE I. COMPARATIVE EXPERIMENT

method	P@1	P@2	P@3
method of	0.412	0.537	0.564
method of Song Wanpeng	0.423	0.573	0.605
method of the paper	0.483	0.581	0.627

V. CONCLUSION

The paper's method represents the semantic information of short text message thoroughly so as to improve the accuracy of similarity calculation. We selected a large number of problem test sets for experiments. This method is feasible and applicable.

REFERENCES

[1] Q.Liu and S.J.li, "Word's semantic similarity computation Based on the HowNet", The 3rd Chinese lexical and semantic proseminar, Taipei, China, 2002.

[2] Tian Weidong, Li Yajuan. center word recognition based on CRF and error driven .[J]. Application Research of computers.2013.

[3] http://www.datatang.com/data/44720

[4] Oliva J. Serrano J I. Del Castillo M D. et al. SyMSS: A syntax-based measure for short-text semantic similarity[J].Data&Knowledge Engineering.2011.70(4): 390-405.

[5] Huang Chenghui. Jian Yin. Hou Fang. A combination text similarity measure method of word semantic information and TF-IDF method [J].Computer science.2011.34 (5): 857-864.

[6] Song W.Liu W.Gu N.A Semantic Space for Question Similarity Calculation in User-Interactive Question Answering Systems. Journal of Computational Information Systems. 5(3): 1055-1063. June. 2009.

[7] B.Ge,F.F.Li,S.L.Guo, "Word's semantic similarity computation method based on HowNet", Application Reserarch of Computers, Vol.27, No.9, pp.3329-3333, Sep.2010.

[8] Hu, Feng Song, Guo, Yong, "An improved algorithm of word similarity computation based on HowNet", Computer Science and Automation Engineering, IEEE International Conference, Vol.3, May 2012.

[9] Z.Dong and Q.Dong,HowNet,http://www.keenage.com.

[10] He X.Liu L.Wu J. Semantic Similarity Calculation Based on Sememe Set. In:Proc of the 2010 International Conference on Artificial Intelligence and Computational Intelligence. Sanya. China: IEEE Computer Society.2010.423-428

[11] Feng song Hu. Yong Guo. An Improved Algorithm of Word Similarity Computation Based on HowNet. In: Proc of the 2th International Conference on Computer Science and Automation Engineering. Zhang jia jie. China. 2012

[12] Luo Jun. Ke Zhang and Xilin Chen . Text Similarity Computing Based on Sememe Vector Space. IEEE ICSESS 2013.